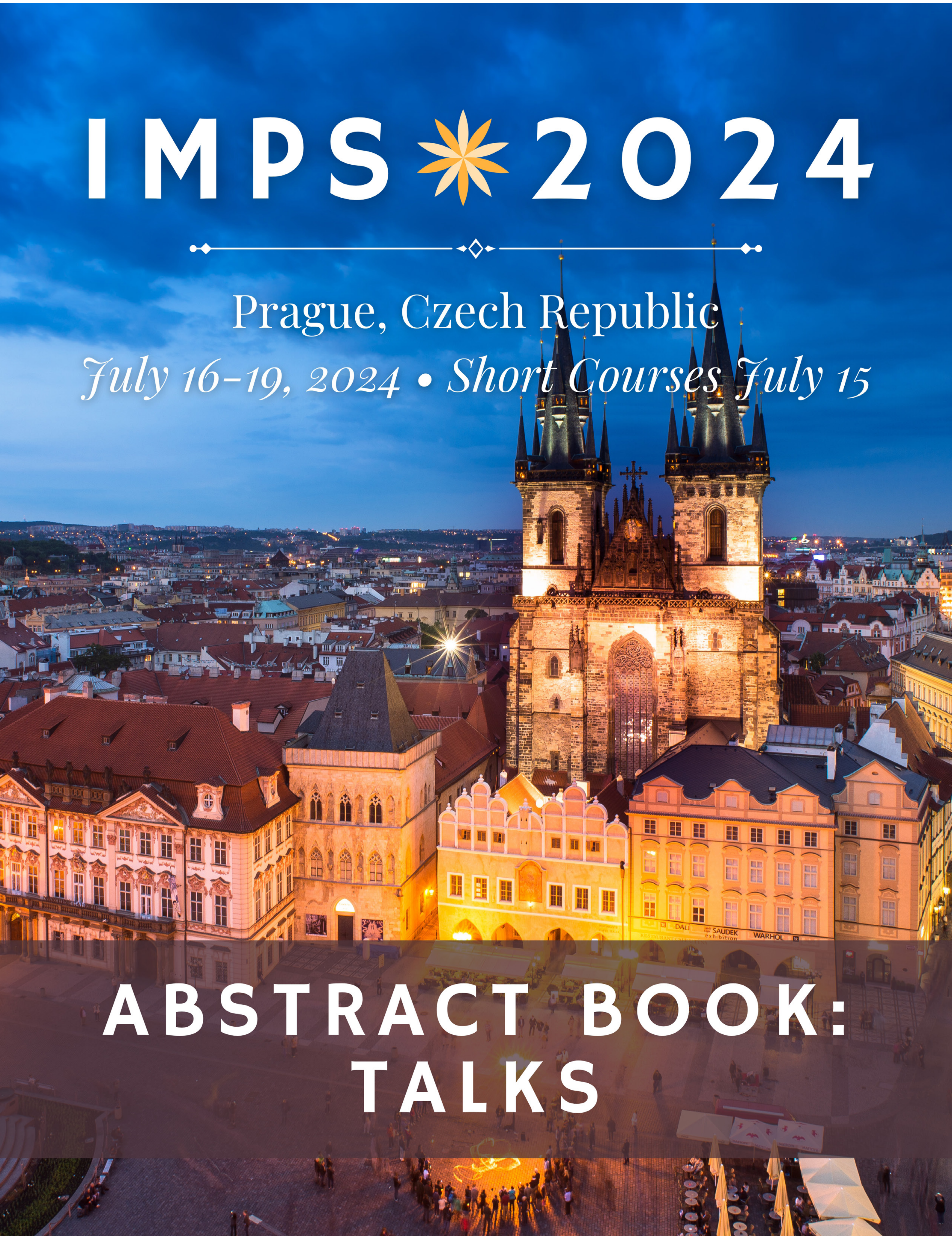
IMPS 2024

Prague, Czech Republic

July 16–19, 2024 • Short Courses July 15

ABSTRACTS

Table of Contents



IMPS 2024

Prague, Czech Republic

July 16-19, 2024 • Short Courses July 15

ABSTRACT BOOK: TALKS

Statistical Inference with Model Uncertainty

Tuesday, 16th July - 09:15: Keynote Address (Vencovského aula) - Keynote Address

Prof. Adrian E. Raftery (University of Washington)

Choosing a statistical model and accounting for uncertainty about this choice are important parts of the scientific process and are required for common statistical tasks such as parameter estimation, interval estimation, statistical inference, point prediction, and interval prediction. A canonical example is the choice of variables in a linear regression model. Many ways of doing this have been proposed, including Bayesian and penalized regression methods, and it is not clear which are best. We compare 21 popular methods via an extensive simulation study based on a wide range of real datasets. We found that three adaptive Bayesian model averaging methods performed best across all the statistical tasks and that two of these were also among the most computationally efficient. We also compared different priors on model space. Finally, we addressed the question of whether model averaging provides an advantage over model selection. This is joint work with Anupreet Porwal.

Quality Standards for Simulation Studies in Psychometrics

Tuesday, 16th July - 10:45: Invited Panel: Quality Standards for Simulation Studies in Psychometrics (Vencovského aula) - Invited Panel

Dr. Terrence Jorgensen (University of Amsterdam), Anna Lohmann (EAH Jena), Björn Siepe (Psychological Methods Lab, Department of Psychology, University of Marburg), Anders Skrondal (University of Oslo), Prof. Carolin Strobl (University of Zurich), Dr. Richard Feinberg (National Board of Medical Examiners)

Simulation studies are an important and often essential tool in methodological research. We use and read about them on an everyday basis, and get excited about, or even despair about, their results. Yet, we tend to share our views, excitement, and frustration only with close peers, if at all. As a scientific community we do not generally discuss what each of us believe makes a good simulation study, and whether or not we agree on these principles and criteria. To stimulate this kind of exchange, this panel discussion will present different views on how simulation studies should be planned, conducted, quality checked, reported, and potentially pre-registered. After brief presentations from the panelists, we will include the audience in an interactive discussion about quality standards for simulation studies in psychometrics.

Bayesian Latent Class Dynamic Mediation Model

Tuesday, 16th July - 10:45: Problems in Mediation Analysis (RB 101) - Oral

Prof. Ying Yuan (University of Texas MD Anderson Cancer Center)

Traditional mediation analysis assumes that a study population is homogeneous and the mediation effect is constant over time, which may not hold in some applications. Motivated by smoking cessation data, we propose a latent class dynamic mediation model that explicitly accounts for the fact that the study population may consist of different subgroups and the mediation effect may vary over time. We use a proportional odds model to accommodate the subject heterogeneities and identify latent subgroups. Conditional on the subgroups, we employ a Bayesian hierarchical nonparametric time-varying coefficient model to capture the time-varying mediation process, while allowing each subgroup to have its individual dynamic mediation process. A simulation study shows that the proposed method has good performance in estimating the mediation effect. We illustrate the proposed methodology by applying it to analyze smoking cessation data.

Relative importance analysis in multiple mediator models

Tuesday, 16th July - 11:00: Problems in Mediation Analysis (RB 101) - Oral

Dr. Xin Gu (East China Normal University), Ms. Xun Zhu (East China Normal University), Mr. Junyu Yan (East China Normal University)

Evaluating relative importance of multiple mediators can help researchers better understand mediation effect and guide interventions. For example, researchers may wish to know whether self-beliefs or learning approaches play a more important role between personality traits and academic performance, or in other words which of them is more worthy of intervention to improve academic performance with limited resources. The traditional coefficient-based measures of indirect effect merely focus on the partial effect of each mediator, which may render undesirable results of importance assessment. This study presents a new method of measuring mediators' relative importance, which attributes the R-squared indirect effect to each mediator using dominance analysis. The new measures can be interpreted as the proportionate contribution of each mediator to the explained variance of indirect effect. Importance ordering of pairwise mediators is inferred using Wald tests and bootstrap confidence intervals and importance ordering of multiple mediators is evaluated using Bayes factors. The proposed method is also extended to latent mediation models. The computation of the importance measures is implemented in an R function that requires only the sample correlation matrix. A real data example is used to illustrate the procedure of assessing the relative importance of mediators.

Assessing heterogeneous treatment effects in Bayesian longitudinal mediation models

Tuesday, 16th July - 11:15: Problems in Mediation Analysis (RB 101) - Oral

Ziwei Zhang (University of Minnesota, Twin-Cities), Dr. Nidhi Kohli (University of Minnesota, Twin-Cities), Dr. Eric Lock (University of Minnesota, Twin-Cities)

Often in educational and psychological randomized trials, researchers discover that sub-groups of individuals respond to the same intervention or treatment differently due to the underlying differences among individuals. Hence, the overall population of individuals constitutes a mixture of two or more unobserved (latent) classes. Previous studies have investigated such heterogeneous treatment effects across different latent classes for both cross-sectional and longitudinal (repeated measures) outcomes. However, it has not been studied in the context of longitudinal mediation analysis. This is a critical gap because longitudinal mediation analysis is frequently utilized for studying substantive research hypothesis. The central focus of such analysis is to estimate the effect of an independent variable (X, i.e., a randomized treatment) on a longitudinal dependent variable (Y, e.g., overall academic performance) via a longitudinal mediator (M, e.g., literacy skills). However, one cannot always assume that X will produce homogenous effects on Y for all participants. Thus, there is a need to develop statistical models within longitudinal mediation analysis to allow for heterogenous effects of X on Y across sub-groups of participants. This leads to the main aim of the study, which is to develop Bayesian longitudinal mediation models to assess heterogeneous treatment effects. The longitudinal variables M and Y could follow intrinsically linear (e.g., linear, quadratic) or nonlinear (e.g., exponential, piecewise) trajectories. Additionally, we enable the model to incorporate class predictive covariates and growth parameter predictive covariates to mimic real data applications. Lastly, we present an empirical data example to illustrate the application of the model.

Causal mediation analysis for binary outcomes with asymmetric link functions

Tuesday, 16th July - 11:30: Problems in Mediation Analysis (RB 101) - Oral

Yuji Tsubota (Osaka University), Dr. Michio Yamamoto (Osaka University / RIKEN AIP)

Causal mediation analysis provides a framework for assessing the extent to which an exposure affects an outcome through an intermediate variable lying between the exposure and the outcome on the causal path. For binary outcomes, traditional methods for causal mediation analysis often use logit and probit link functions that implicitly assume the symmetry of the success probability curves, which may lead to poor performance when this symmetry assumption is badly violated. To address this issue, we utilize the complementary log-log link function which can model the outcome success probability asymmetrically in the development of an alternative regression-based causal mediation analysis framework. We give the definitions of the controlled direct effect, natural direct effect, natural indirect effect in our framework with the complementary log-log link function and explain the conditions for their identification. We derive the closed-form analytic expressions of our causal effects allowing us to estimate the effects by simple regression analyses and leading to clearer interpretation of the effects. To illustrate the efficacy of our proposed methodology, we apply it to the real-world data from educational research for which the complementary log-log model shows a better fit than logit and probit models. We also assess the accuracy of our estimators through numerical simulations.

Symposium on Asymmetric Item Response Models

Tuesday, 16th July - 10:45: Symposium: Asymmetric Item Response Models (NB A) - Symposia

Dr. Jay Verkuilen (City University of New York)

While asymmetric item response theory models (AsymIRT) date back to work by Goldstein (1980) using the complementary log-log link as an alternative to the Rasch model and Samejima (2000) using a Type I generalized logistic distribution as an extension of the 2PL, these models have received a lot of attention from researchers in the last decade. This symposium will cover novel contributions to the literature on AsymIRT. We anticipate a mixture of theory and applications in our papers. There will be four papers on the panel with Dr. Carl Falk acting as a discussant.
Papers are by
Verkuilen & Johnson
Setti & Feuerstahler
Bonifay & Suh
Bolt

Heywood Cases in Asymmetric IRT

Tuesday, 16th July - 10:45: Symposium: Asymmetric Item Response Models (NB A) - Symposia

Dr. Jay Verkuilen (City University of New York), Mr. Peter Johnson (City University of New York)

Heywood cases and other improper solutions occur frequently in latent variable models. They have important consequences for scoring with the latent variable model and are indicative of issues in a model such as poor identification or model misspecification. In the context of the 2PL and 3PL models in IRT, they are more frequently known as Guttman items and identified by a discrimination parameter that is deemed excessively large. Asymmetric item response theory (AsymIRT) models often have parameters that are not easy to interpret directly and so scanning parameter estimates are not necessarily indicative of the presence of problematic values. Graphical examination of the IRF is useful as well, but necessarily subjective and highly dependent on choices of graphical defaults. We propose using the derivatives of the IRF and the item Fisher information functions to bypass the parameters. We illustrate the approach using several AsymIRT models and an empirical example.

Bayesian model averaging of (a)symmetric IRT models in small samples

Tuesday, 16th July - 10:45: Symposium: Asymmetric Item Response Models (NB A) - Symposia

Fabio Setti (Fordham University), Dr. Leah Feuerstahler (Fordham University)

Symmetric models have long been dominant in item response theory (IRT). However, asymmetric IRT models have recently experienced a surge in popularity due to their ability to reflect complex cognitive processes and to accommodate skewed latent trait distributions. The current simulation study proposes Bayesian model averaging (BMA) as a method of flexibly estimating item response functions (IRFs) in small samples. In addition to the well established Rasch and 2PL models, recently proposed asymmetric models allow for more varied IRF shapes with only one or two parameters per item. This simulation study compares how model selection (MS), BMA, and kernel smoothing IRT (KS) recover IRFs under complex data generating conditions and small sample sizes (i.e., 100 and 250). Models were estimated using the brms package and both stacking weights and BMA weights with bootstrapping were derived from leave-one-out cross-validation. The results showed that BMA consistently outperformed MS and KS under the vast majority of the simulated conditions. The methods proposed in this study may also provide ways of accommodating differences in the scale of the latent trait implied by different IRT models. We believe that BMA using symmetric and asymmetric models offers stable yet flexible model estimation and a compelling alternative to other semi-parametric IRT methods.

Parsimonious item response theory modeling with different link functions

Tuesday, 16th July - 10:45: Symposium: Asymmetric Item Response Models (NB A) - Symposia

Dr. Hyejin Shim (University of Missouri), Dr. Wes Bonifay (University of Missouri)

Traditional item response theory (IRT) models assume a symmetric error distribution and rely on symmetric (logit or probit) link functions to model the response probabilities. However, the symmetry assumption does not always hold in item response data. To explore the benefits of alternate link functions, we investigated a set of one-parameter IRT models for unidimensional tests with dichotomous items by specifying the complementary log-log (CLL), negative log-log model (NLL), and cauchit links. In a series of simulations studies, we demonstrated that these alternate-link models are far more (parametrically) parsimonious than more traditional models, yet comparable with regard to (1) data distribution shape, (b) inflection point shift, (c) structural similarities, and (d) response behaviors. Importantly, the simpler alternate-link models can often outperform more complex traditional models, and these four properties help us to explain why.

Specifically, we demonstrate that the CLL model accounts for guessing effects because it has an item characteristic curve with a higher inflection point than that of the Rasch or two-parameter logistic (2PL) models. Similarly, the NLL model accounts for slipping effects via a lower inflection point. We also present the cauchit model, which addresses both guessing and slipping effects. Our simulations reveal that these one-parameter alternate-IRT models are robust to small sample sizes (e.g., N = 100) and facilitate item-weighted scoring. We conclude by applying the CLL, NLL, and cauchit models to empirical data, thereby providing further evidence for our simulation results.

Unipolar IRT and vocabulary development

Tuesday, 16th July - 10:45: Symposium: Asymmetric Item Response Models (NB A) - Symposia

Prof. Daniel Bolt (University of Wisconsin - Madison), Qi Huang (University of Wisconsin - Madison), Xiangyi Liao (University of Wisconsin - Madison)

We review possible explanations for systematic positive asymmetry in item characteristic curves, including associations between the level of a proficiency and its diversification. Using vocabulary development as an example, we present arguments for a unipolar IRT representation of vocabulary proficiency, and examine how such a representation can explain empirical inconsistencies in the observations of Matthew/anti-Matthew effects in the development of reading and vocabulary.

Uncertainty Quantification for Latent Variable Selection in High Dimensional Exploratory Factor Analysis.

Tuesday, 16th July - 10:45: Topics in Latent Variable Estimation (NB B) - Oral

Ms. Xinyi Liu (The London School of Economics and Political Science), Dr. Yunxiao Chen (The London School of Economics and Political Science), Prof. Irini Moustaki (The London School of Economics and Political Science)

We propose a method for selecting a sparse loading matrix that uses a non-smooth rotation criterion, such as the Lp rotation. This method makes it easier to interpret the factors. However, because of the non-smoothness of the objective function, the central limit theory is invalid for the resulting rotated loading matrix. This makes it difficult to quantify the uncertainty with latent variable selection.

To address this issue, we propose a new approach. We use the non-smooth rotation criterion to identify partial loading structures and then follow this with a confirmatory factor analysis (CFA). Our method establishes an asymptotic central limit theory applicable as the number of items and individuals tend to infinity.

Furthermore, we construct mirror statistics based on the central limit theorem. These statistics effectively allow us to control the false discovery rate in latent variable selection. We conducted a series of experiments to validate our inference results and the effectiveness of our variable selection methods. The empirical evidence from these experiments supports our proposed method.

Pairwise Stochastic Approximation for Confirmatory Factor Analysis of Categorical Data

Tuesday, 16th July - 11:00: Topics in Latent Variable Estimation (NB B) - Oral

Dr. Giuseppe Alfonzetti (University of Udine), Prof. Ruggero Bellio (University of Udine), Dr. Yunxiao Chen (The London School of Economics and Political Science), Prof. Irini Moustaki (The London School of Economics and Political Science)

A widely spread approach for estimating latent variable models of categorical items is the pairwise likelihood method, a limited-information strategy relying on surrogate likelihoods built by stacking together all possible bivariate margins of the data (see Varin et al., 2011 and Katsikatsou et al., 2012). In contrast with full-information maximum likelihood, a pairwise likelihood avoids the need to evaluate high dimensional integrals. Nevertheless, it can still be demanding for large-scale problems involving many observed items because of the high number of possible pairs involved. To unlock the viability of pairwise likelihood estimation on larger datasets, we tackle such computational bottleneck by proposing an approximate estimator derived from an optimisation routine based on stochastic gradients.

While the stochastic optimisation literature typically considers a stochastic gradient as the log-likelihood score evaluated on a small subset of the sample (see Toulis & Airoldi and references therein), our proposal relies on stochastic gradients built by subsampling the possible bivariate margins of the data instead of the observational units. We show that such a procedure leads to a computationally scalable stochastic estimator that is asymptotically equivalent to the pairwise likelihood. Additionally, we highlight that finite sample performances improve when compounding the sampling variability of the data with the uncertainty introduced by the stochastic gradients.

The proposed method is validated with both simulated experiments and real data.

Generalized linear latent variable models with nonignorable missing data

Tuesday, 16th July - 11:15: Topics in Latent Variable Estimation (NB B) - Oral

Dr. Björn Andersson (Centre for Educational Measurement (CEMO), University of Oslo)

Generalized linear latent variable models are typically estimated with likelihood-based approaches that assume missing at random for any missing data. However, in many cases this assumption is not tenable such as when grades of elective subjects are missing or for item nonresponse in a scale or test. We define generalized linear latent variable models that incorporate nonignorable missing data for mixed observed variables, including support for joint modelling of categorical, count, and continuous data. The procedure uses an established approach wherein the observed data are augmented with additional variables defined by the presence or absence of missing data for some variables. We derive a marginal maximum likelihood estimation method for such models that uses adaptive quadrature or Laplace approximations and investigate its properties in simulations. The method is applied to multidimensional item response modelling of grades from elective school subjects in Norway, where nonignorable selection effects are present. We discuss further applications of the approach to account for effects of item nonresponse in large-scale assessments and surveys.

A Two-Stage Path Analysis Approach to Model Interaction Effects for Congeneric Measures

Tuesday, 16th July - 11:30: Topics in Latent Variable Estimation (NB B) - Oral

Mr. Gengrui Zhang (University of Southern California), Dr. Hok Chio (Mark) Lai (University of Southern California)

Interaction effects among latent variables become increasingly popular in psychology research with in-depth theory and complex data structure. Compared to widely used methods of modeling latent interactions, Unconstrained Product Indicator (UPI; Marsh et al., 2004) and Reliability-Adjusted Product Indicator (RAPI; Hsiao et al., 2018), an extended model based on the two-stage path analysis (2S-PA) framework, namely 2S-PA-Int, was evaluated. 2S-PA-Int uses factor scores estimated from congeneric items as single indicators of latent variables and applies error-variance constraints on the single indicators to account for measurement error. Thus the model specification of 2S-PA-Int is simpler than UPI which uses multiple product indicators. We conducted a simulation study with manipulated sample size, covariance between first-order latent predictors, and reliability of congeneric measures across 2000 replications. The results showed that 2S-PA-Int consistently produced the interaction estimates with less standardized bias, acceptable relative SE bias and coverage rates, and lower RMSE values than UPI and RAPI, particularly under the conditions of small sample size and low level of reliability. Technical details of 2S-PA-Int will be discussed. In the second study, we provided step-by-step demonstrations of applying the three methods on an empirical dataset sourced from the Panel Study of Income Dynamics (PSID). We showed that the interaction estimates using the empirical data were consistent with our simulation results. Overall, we argued that 2S-PA-Int is able to accommodate more extreme conditions with a small sample size and low reliability, while maintaining model simplicity and reducing risks of non-convergence.

Ability Estimation for Culturally Responsive Assessments

Tuesday, 16th July - 11:45: Topics in Latent Variable Estimation (NB B) - Oral

Dr. Sandip Sinharay (Educational Testing Service)

- In a NAEP validity studies panel commissioned report, Hughes (2023) recommended widespread use of socioculturally responsive assessments (SCRA) acknowledging the influence of sociocultural factors on test performance, and additional research to estimate the sizes of the effects of different sociocultural factors on the test performance of racially and culturally diverse groups of test takers. We will discuss these recommendations and the results from a small study regarding SCRAs.

Integrating Large Language Models and Artificial Intelligence in Research: Network Psychometrics, Emotion Dynamics, and Open-Ended Evaluations

Tuesday, 16th July - 10:45: Symposium: Integrating Large Language Models and Artificial Intelligence in Research: Network Psychometrics, Emotion Dynamics, and Open-Ended Evaluations (NB C) - Symposium Overview

Dr. Hudson Golino (University of Virginia)

This symposium aims to present new applications of large language models and artificial intelligence techniques in psychological research, from network psychometrics to emotion dynamics and the evaluation of open-ended responses. Lara Russell-Lasalandra will present her work on ”**Assessing the Quality of AI-Generated Items: A Network Psychometric Approach**“. Dr. Aleksandar Tomašević will present the work on ”**Decoding emotion dynamics using transformer-based AI models: A novel approach for facial expression recognition in videos using transforEmotion R package**“. Dr. Hudson Golino will talk about ”**Improving Retrieval-Augemented Generation and Zero-Shot Classification with Exploratory Graph Analysis**“. Dr. Mariana Teles will present her work on ”**Extracting useful information from Open-Ended Responses using Zero-Shot Classification: an example using student’s evaluation of teaching**”.

Assessing the Quality of AI-Generated Items: A Network Psychometric Approach

Tuesday, 16th July - 10:45: Symposium: Integrating Large Language Models and Artificial Intelligence in Research: Network Psychometrics, Emotion Dynamics, and Open-Ended Evaluations (NB C) - Symposia

Mrs. Lara Russell-Lasalandra (University of Virginia), Dr. Hudson Golino (University of Virginia)

The increasing ubiquity and efficacy of Large Language Models (LLM) has empowered researchers to explore once implausible projects. One such project might be novel item generation. Drafting several hundred items has the potential to be incredibly resource intensive. Therefore, researchers may be tempted to use an LLM to automatically generate as many items as needed. The present research introduces a pipeline that can assess the quality of the AI-generated items using exploratory graph analysis. The pipeline was tested using GPT-generated items that mimic those on the Big Five personality assessment. While the use of LLMs in psychological scale development can make the task significantly less daunting, some computer-authored items do not approximate the quality of those that are expert-authored. The present pipeline allows researchers to remove suboptimal items so they can be more confident when testing and ultimately implementing scales that contain AI-generated items.

Decoding emotion dynamics using transformer-based AI models: A novel approach for facial expression recognition in videos using transforEmotion R package

Tuesday, 16th July - 10:45: Symposium: Integrating Large Language Models and Artificial Intelligence in Research: Network Psychometrics, Emotion Dynamics, and Open-Ended Evaluations (NB C) - Symposia

Dr. Aleksandar Tomašević (University of Novi Sad), Dr. Hudson Golino (University of Virginia), Dr. Alexander Christensen (Vanderbilt University)

In this work, we propose a novel machine-learning approach for detecting emotions from facial expressions in videos. We leverage recent advancements in deep learning and transformer-based neural network architectures to extract time series analysis of Facial Expression Recognition (FER) scores. Our approach is implemented in the transforEmotion R package, which provides easy access to transformer-based models available via the HuggingFace service. Notably, this R package enables zero-shot emotion classification of text, images, and videos in R, without the need for a GPU, subscriptions, paid services, or knowledge of Python.

The data provided by this package is valuable for psychological research on affective dynamics and emotion expression dynamics. However, the properties of such data are not well understood in the current literature. To address this gap, we conduct a large-scale simulation and employ dynamic Exploratory Graph Analysis (DynEGA) to investigate the dimensionality of the data and uncover the latent dimensions underlying FER scores, specifically the positive and negative sentiment of expressed emotions during a public speech. Our results demonstrate that DynEGA successfully reveals the latent structure of emotion dynamics expressed through FER scores, both in generative simulation and real video footage analyzed using the transforEmotion package.

In conclusion, our research contributes to the understanding of emotion dynamics in facial expressions by proposing a novel approach for modeling FER scores. The application of our methods can benefit psychological research and contribute to the development of more accurate emotion recognition systems.

Improving Retrieval-Augemented Generation and Zero-Shot Classification with Exploratory Graph Analysis

Tuesday, 16th July - 10:45: Symposium: Integrating Large Language Models and Artificial Intelligence in Research: Network Psychometrics, Emotion Dynamics, and Open-Ended Evaluations (NB C) - Symposia

Dr. Hudson Golino (University of Virginia)

Retrieval-Augmented Generation (RAG) and Zero-Shot classification (ZSC) are two approaches within Large Language Models that can provide additional context to a model and generate scores for unseen categories during training. In this presentation, I will show how network psychometrics and exploratory graph analysis can be used to improve the quality of an RAG+ZSC for language tasks.

Extracting useful information from Open-Ended Responses using Zero-Shot Classification: an example using student's evaluation of teaching

Tuesday, 16th July - 10:45: Symposium: Integrating Large Language Models and Artificial Intelligence in Research: Network Psychometrics, Emotion Dynamics, and Open-Ended Evaluations (NB C) - Symposia

Dr. Mariana Teles (University of Virginia)

This study explores the application of Zero-Shot Classification, utilizing Facebook's BART Large Language Model, to extract meaningful insights from students' open-ended responses in course evaluations. A comparison was conducted between two iterations of an Introduction to Cognition course. The traditional lecture-based format offered in 2022 was contrasted with a student-centered active learning approach implemented in Spring 2023, with feedback gathered from 198 participating students. This research aimed to discern the pedagogical impact of these differing methodologies on the student learning experience through advanced Natural Language Processing (NLP) techniques.

Utilizing the transforEmotion package in R for analysis, the study employed Zero-Shot Classification to categorize responses into four distinct areas: Better Learning Experience, Learned More, Engaged More, and More Excited About the Content. Zero-shot classification enables the categorization of text into previously unseen categories, thereby facilitating the analysis of open-ended feedback without the constraints of pre-labeled data sets. This approach highlighted significant enhancements in the active learning course across three categories: Better Learning Experience, Engaged More, and Learned More when compared to the traditional lecture-based format. However, excitement about the content showed no significant difference between the two pedagogical approaches.

These findings underscore the potential of machine learning and NLP in educational research, particularly in evaluating and improving teaching methodologies. By leveraging Zero-Shot Classification, educators and researchers can gain deeper insights into student feedback, fostering a more informed approach to pedagogical innovation.

Maximal Reliability Coefficients for Bifactor Models

Tuesday, 16th July - 10:45: Topics in Reliability (NB D) - Oral

Victoria Savalei (University of British Columbia), Sijia Li (University of British Columbia)

Reliability coefficients are frequently used to evaluate solution quality in an orthogonal bifactor model. Methodologists recommend both coefficients omega (which give reliabilities of unweighted composites) and maximal reliability coefficients (which give reliabilities of optimally weighted composites) for this purpose. In the context of the 1-factor model, maximal reliability coefficient has also been rebranded as “Coefficient H” (Hancock, 2001). Despite its derivation under the 1-factor model and not the bifactor model, this coefficient has been widely recommended and used to assess the quality of measurement of the group factors in bifactor models as well. We will show that the 1-factor model equation for coefficient H does not correspond to the reliability of any composite under a bifactor model. We will also provide equations for the correct version of maximal reliability coefficients under a bifactor model. As in the case of the 1-factor model, these coefficients turn out to be equivalent to the measures of factor determinacy (FD; Rodriguez et al., 2016a). Lastly, we will show that in the case of orthogonal bifactor models (unlike in the case of orthogonal confirmatory factor models), maximal reliability coefficients based on the entire model assign nonzero weights to all variables in the model and not just to the indicators of the group factor being assessed, complicating their interpretation. We provide a simplified version of maximal reliability coefficients for bifactor models that set weights of irrelevant indicators to zero and we evaluate the performance of the simplified versions in a simulation.

On the estimation of reliability in knowledge space theory

Tuesday, 16th July - 11:00: Topics in Reliability (NB D) - Oral

Prof. Andrea Spoto (University of Padua), Dr. Debora de Chiusole (University of Padua), Dr. Umberto Granziol (University of Padua), Prof. Luca Stefanutti (University of Padua)

Unlike classical test theory, where the overall reliability of a test is measured based on the ratio between true score variance and total variance, in knowledge space theory (KST), to date, there are no indices for assessing the overall reliability of a test. The only available methods focus on specific items, rather than on the test as a whole. Moreover, the classical indices of reliability cannot be used in KST due to the non-independence of error and true score in this framework. In this presentation two indices for the overall assessment of test reliability in KST are introduced, based on the concepts of entropy and conditional entropy. More precisely, the Response Pattern Reliability index (RP-Reliability) is designed to evaluate the reliability of response patterns in relation to the knowledge state, while the Knowledge State Reliability index (KS-Reliability) is designed to assess the reliability of an individual's estimated knowledge state. The presentation will offer theoretical insights and some simulations to explore and test the performance of the indices under various conditions. Additionally, the presentation will focus on how these approaches may enhance test reliability assessment within the context of KST, complementing other available methods.

Reliability in intensive longitudinal studies: A state-space vs. linear mixed model approach

Tuesday, 16th July - 11:15: Topics in Reliability (NB D) - Oral

Mr. Tzu-Yao Lin (Maastricht University), Prof. Francis Tuerlinckx (KU Leuven), Prof. Sophie Vanbelle (Maastricht University)

With the advancement of technology, measures are more and more often collected in real-time resulting in intensive longitudinal data (ILD). ILD allow us to study dynamic properties of one or more variables, such as moment-to-moment fluctuations in affect or changes in heartbeat rates. Typically, for ILD, the fact is that observations with a person are serially correlated (sometimes at a high frequency). Despite the widespread use of ILD, the issue of measurement error and reliability in the presence of serial correlation has not been studied extensively. Some notable exceptions are Schuurman and Hamaker (2019), Vangeneugden et al. (2004), and Laenen et al. (2007). The former relies on a Bayesian state space model, while the latter two consider a linear mixed model framework. In this talk, we will compare both frameworks and elucidate similarities and differences regarding the issue of measurement error and reliability. To illustrate both approaches, we have re-analyzed data on positive and negative affects, where individuals were measured multiple times a day for several days. We also discuss the within-person and between-person reliability, which can vary over time.

Estimating Reliability for Mixed-Format Tests

Tuesday, 16th July - 11:30: Topics in Reliability (NB D) - Oral

Ye Yuan (University of Georgia), Ying Lu (College Board), Amy Hendrickson (College Board)

The purpose of the research is to compare the performance of different reliability estimation approaches and provide recommendations for reliability estimation for a test program with mixed-format tests. Estimates of reliability based on classical internal consistency and item response theory (IRT) reliability estimation procedures are examined for five mixed-format tests.

First, the split-half reliability estimates under the assumptions of classically parallel part-tests (Spearman-Brown prophecy formula) and under the assumptions of essentially tau-equivalent part-tests (Flanagan, 1937) are examined. Second, the reliability of the composite score as a linear composite of multiple-choice (MC) section score and free-response (FR) section score is calculated. Third, treating each item as a part-test, Alpha, Raju, Feldt and Feldt-Gilmer coefficients are calculated. Lastly, a three-parameter logistic model is fit to the MC items and a general partial credit model is fit to the FR items. After calculating the true score variance and the raw score error variance across all persons, the research uses model-based estimation of observed score variance to calculate the IRT marginal reliability.

Our study found that it is not viable to directly calculate Coefficient alpha for mixed-format tests. Coefficient alpha is negatively biased by the use of parts that are not essentially tau-equivalent. All the stratified approaches produce similar results. If there is concern that the FR part length is not a good indicator of effective length, the classical congeneric instead of Raju coefficient can be used. Moreover, the IRT approach provides estimates very close to split-half, stratified approaches and Feldt/Feldt-Gilmer coefficients.

Standard errors and null-hypothesis significance tests for reliability coefficients

Tuesday, 16th July - 11:45: Topics in Reliability (NB D) - Oral

Andries van der Ark (University of Amsterdam)

Reliability analysis is one of the most conducted analyses in applied psychometrics. It entails the assessment of reliability of both item scores and scale scores using coefficients that estimate the reliability (e.g. Cronbach's alpha), estimate measurement precision (e.g., estimated standard error of measurement), or estimate the contribution of individual items to the reliability (e.g., corrected item-total correlations). Most statistical software packages used in the social and behavioral sciences offer these reliability coefficients. Standard errors and null-hypothesis significance tests (NHSTs) are generally unavailable for reliability coefficients, which is a bit ironic for coefficients that are about measurement precision. For a large number of coefficients, I derived standard errors and NHSTs. In this presentation, I will discuss the dilemmas and challenges of this task. In particular, I will discuss (1) categorical marginal models (CMMs), which I used as a framework for finding the correct sampling distributions of reliability statistics, (2) the challenges of estimating CMMs when a large number of items is involved, and (3) the challenges of developing user-friendly R software. Finally, I will show what I think reliability-analysis computer output should look like in the future.

Advancements in Structural Equation Modeling (SEM): Model Fit, Parameter Selection, and Measurement Invariance

Tuesday, 16th July - 10:45: Symposium: Advancements in Structural Equation Modeling (SEM): Model Fit, Parameter Selection, and Measurement Invariance (RB 209) - Symposium Overview

David Goretzko (Utrecht University)

Within the social sciences, the measurement of unobservable phenomena through observable indicators is a central component of statistical analyses. Ensuring the reliability and validity of measurement models across populations is essential for obtaining accurate results. This symposium explores both the limitations of conventional methods and benefits of new statistical approaches in evaluating model fit, selecting parameters or variables, and testing measurement invariance.

In the first talk, Melanie Partsch examines the sensitivity of model fit indices in SEM to misspecifications within the measurement model, particularly focusing on underfactoring (i.e., specifying an insufficient number of latent variables).

Next, Sara van Erp introduces projection predictive variable selection as a novel method to select parameters within the SEM framework. This approach aims to improve upon parameter selection in Bayesian regularized SEM based on arbitrary cutoff values for the estimates or decision rules based on a 95% credibility interval.

David Goretzko explores the application of model-based recursive partitioning for analyzing measurement invariance. He introduces recently developed techniques including EFA trees and their network-based counterpart, EGA trees.

Ai Ye focuses on the impact of individual differences in the within-person measurement model (i.e., between-person measurement non-invariance using intensive longitudinal data) on model selection and estimation in dynamic factor models.

Lastly, Philipp Sterner conceptualizes measurement invariance as a causal inference problem utilizing directed acyclic graphs (DAGs). He explains how this perspective on non-invariance can advance our theoretical knowledge about the measurement process and aid in selecting suitable modeling approaches.

Sensitivity of model fit indices in SEM to underfactoring

Tuesday, 16th July - 10:45: Symposium: Advancements in Structural Equation Modeling (SEM): Model Fit, Parameter Selection, and Measurement Invariance (RB 209) - Symposia

Dr. Melanie Viola Partsch (Utrecht University), Philipp Sterner (Ludwig-Maximilians-Universität München), David Goretzko (Utrecht University)

When evaluating the fit of their SEM, researchers often rely on goodness-of-fit indices and associated cut-off rules, such as Hu and Bentler's (1999) recommendation to accept a model if CFI ≥ .95, RMSEA ≤ .06, and SRMR ≤ .08. However, different fit indices are sensitive to different types of misspecifications, and they all depend not only on model (mis-)fit but also on various nuisance parameters, such as sample size. Furthermore, cut-off recommendations are not simply generalizable to models, data conditions, and misspecifications neglected in the simulation studies they were derived from. We present a study that focuses on the sensitivity of commonly used model fit indices to *underfactoring*, that is, the specification of an insufficient number of latent variables in a CFA measurement model—a severe misspecification whose impact on model fit has not yet been sufficiently investigated and that was usually not considered in simulation studies from which cut-off recommendations were derived. Based on 432,000 datasets from a broad simulation study, we examine how the most commonly used fit indices CFI, RMSEA, and SRMR behave if underfactored vs. correctly specified models have been fitted. Thereby, we take into account the impact of various nuisance parameters, such as sample size, the number of latent and manifest variables, loading sizes, and factor correlation patterns, and their interaction with model (mis-)specification. Against the background of the results, we discuss the caveats of applying CFI, RMSEA, SRMR, and their common cut-offs for model fit evaluation in the presence of underfactoring.

Projection predictive variable selection for Bayesian regularized structural equation modeling

Tuesday, 16th July - 10:45: Symposium: Advancements in Structural Equation Modeling (SEM): Model Fit, Parameter Selection, and Measurement Invariance (RB 209) - Symposia

Dr. Sara van Erp (Utrecht University), Prof. Paul-Christian Bürkner (TU Dortmund University), Prof. Aki Vehtari (Aalto University)

Classical regularized structural equation modeling (SEM) relies on optimization with a penalty function added to the usual estimation problem. An alternative to the classical approach is Bayesian regularized SEM in which the prior distribution serves as penalty function. Many different shrinkage priors exist, enabling great flexibility in terms of shrinkage behavior. Additionally, advantages in terms of automatic uncertainty estimates, the possibility to include prior knowledge, and intuitive interpretation of the results have resulted in various applications of Bayesian regularized SEM.

The goal of Bayesian regularized SEM is often to select a more parsimonious model by including only those parameters in the model which show substantial effects after regularization. Currently, ad-hoc methods are used in SEM to decide if a parameter estimate should be set to zero or not for example by relying on an arbitrary threshold value or on the 95% credibility interval. However, it has been shown that the optimal selection criterion depends on various sample and model characteristics. Thus, a formal selection method that works well across different types of SEMs and conditions is needed.

A promising method that is available in regression models is projection predictive variable selection, which offers a practical approach of selecting the model that offers nearly similar predictions as a reference model. In this presentation, I will present an extension of the projection predictive method to SEM to determine which parameter estimates to set to zero, thereby performing automatic model selection.

Exploring Measurement Non-Invariance with Exploratory Factor Analysis Trees and Exploratory Graph Analysis Trees

Tuesday, 16th July - 10:45: Symposium: Advancements in Structural Equation Modeling (SEM): Model Fit, Parameter Selection, and Measurement Invariance (RB 209) - Symposia

David Goretzko (Utrecht University), Philipp Sterner (Ludwig-Maximilians-Universität München)

Psychological research frequently deals with unobservable constructs like cognitive abilities or personality traits, which are inferred from scales and test items designed to measure them. When comparing latent variables representing these constructs across groups, ensuring measurement invariance (MI) is crucial. However, standard procedures for testing MI have limitations and may not adequately facilitate comprehensive exploration of MI.

Model-based recursive partitioning (MOB) provides a framework to create statistical approaches that can address MI at the earliest stages of scale development taking into account numerous covariates. In this presentation, we introduce two innovative approaches that merge MOB with highly exploratory methodologies. The first approach, known as EFA trees, utilizes Exploratory Factor Analysis (EFA) to examine MI across several covariates and diverse sub-populations where the measurement model may vary substantially. While EFA trees offer unparalleled exploratory potential compared to traditional methods for MI analysis, detecting severe violations of configural invariance, stemming from differing numbers of latent factors across populations, requires combining EFA trees with a suitable factor retention criterion. To address this, we have replaced EFA with Exploratory Graph Analysis (EGA) to develop EGA trees, enabling a more thorough exploration of subgroups exhibiting divergent factor solutions.

Measurement Invariance in Within-Person Latent Structure for Dynamic Factor Models

Tuesday, 16th July - 10:45: Symposium: Advancements in Structural Equation Modeling (SEM): Model Fit, Parameter Selection, and Measurement Invariance (RB 209) - Symposia

Dr. Ai Ye (Ludwig-Maximilians-Universität München)

The dynamic factor model (DFM; Browne & Nesselroade, 2005; Molenaar, 1985) is commonly used to model multivariate time series data with measurement error. Model selection and estimation methods for DFM are investigated by many researchers. Under idiographic approaches, heterogeneous structural models are used to maximize individual differences in their dynamics (Ye & Bollen, 2022). However, it is always assumed that the measurement model is homogeneous across people, i.e., between-person measurement invariance is assumed, for model identification and interpretation. In this novel study, I examined the same issues of model selection and estimation for DFM with the flexibility of having heterogeneous measurement models, i.e., between-person measurement non-invariance in their within-person longitudinal latent structure. I used the LASSO regularization under the Structural Equation Model framework to select the nonzero paths from a semi-confirmatory measurement model. I investigate the sensitivity to recover true paths and the specificity to eliminate false ones in the data-generating sparse model. The other goal is to obtain robust estimates using the model-implied instrumental variable, two-stage least squares (MIIV-2SLS; Bollen, 1996) estimation, especially that the model selection procedure introduces misspecification in the measurement at an individual level. The proposed method highlights the flexibility in modeling individual DFMs with heterogeneous measurement models, with a robust estimation. We aim to offer researchers guidance on model selection and estimation in person-centered dynamic assessments.

A causal framework for the comparability of latent variables

Tuesday, 16th July - 10:45: Symposium: Advancements in Structural Equation Modeling (SEM): Model Fit, Parameter Selection, and Measurement Invariance (RB 209) - Symposia

Philipp Sterner (LMU Munich), Florian Pargent (LMU Munich), Dominik Deffner (Max Planck Institute for Human Development Berlin), David Goretzko (Utrecht University)

Measurement invariance (MI) describes the equivalence of measurement models of a construct across groups or time. When comparing latent means, MI is often stated as a prerequisite of meaningful group comparisons. The most common way to investigate MI is multi-group confirmatory factor analysis (MG-CFA). Although numerous guides exist, a recent review showed that MI is rarely investigated in practice. We argue that one reason might be that the results of MG-CFA are uninformative as to why MI does not hold between groups. Consequently, under this framework, it is difficult to regard the study of MI an interesting and constructive step in the modeling process. We show how directed acyclic graphs (DAGs) from the causal inference literature can guide researchers in reasoning about the causes of non-invariance. For this, we first show how DAGs for measurement models can be translated into the path diagrams used in the linear structural equation model (SEM) literature. We then demonstrate how insights gained from this causal perspective can be used to explicitly model encoded causal assumptions with moderated SEMs, allowing for a more enlightening investigation of MI. Ultimately, our goal is to provide a framework in which the investigation of MI is not deemed a "gateway test" that simply licenses further analyses. By enabling researchers to consider MI as an interesting part of the modeling process, we hope to increase the prevalence of investigations of MI altogether.

A New Measure and an Old Measure of Person Fit

Tuesday, 16th July - 10:45: Developments in IRT (RB 210) - Oral

Mr. Horacio Rocha (CUNY Graduate Center), Dr. Jay Verkuilen (CUNY Graduate Center)

Most measures of person fit are highly correlated to Guttman errors, these include Ht, Lz and Zh statistics. This paper examines new non-parametric measures of person fit that have less correlation to Guttman errors and as a result can be additive when used in conjuntion with Guttman errors. One meausre is the sumproduct (SP statistic) of the respondent vector and the proportion of correct response. In a simulation study, 10% errors were added to the SAPA dataset, the Guttman errors correctly identified 48% of the simulated responses, while Ht identified 65% of the responses and the SP measure identified 45% of responses. Thought Guttman and SP indentified a smaller percent than Ht, the correlation of responses between Ht and Guttman was high, while the correlation of indentified responses between Guttman and SP was very low. Guttman identified 90% of the responses identifed by Ht, while SP and Guttman only had 23% similar identification. The full paper will include several different types f simulations and will include two additional non-parametric person fit measures.

Detection of local dependence within PCMs through recent L1-penalization approaches

Tuesday, 16th July - 11:00: Developments in IRT (RB 210) - Oral

Mr. Can Gürer (UMIT TIROL Private University for Health Sciences and Technology), Dr. Clemens Draxler (UMIT TIROL Private University for Health Sciences and Technology)

The presence of local dependence in Item Response Theory (IRT) models can lead to biased ability and item parameter estimates and numerous techniques have been presented for its detection within the binary Rasch Model. Yet the question of local dependence within the Partial Credit Model has not been touched upon extensively. This talk aims to assess the viability and performance of modern linear penalization techniques within IRT especially for polytomous items, i.e. the cmlPCMlasso or the GPCMlasso. The performance of these approaches regarding detection is illustrated through examples and simulations. Additionally, an approach to managing local dependence based on the findings of Verhelst & Verstralen (2008) is provided.

Investigating the Invariant Ordering of Clustered Items Using Nonparametric IRT

Tuesday, 16th July - 11:15: Developments in IRT (RB 210) - Oral

Dr. Letty Koopman (University of Groningen), Prof. Johan Braeken (Centre for Educational Measurement (CEMO), University of Oslo)

Invariant item ordering (IIO) is a property of a test or questionnaire, in which the items have the same order in difficulty (or popularity) for all values of a unidimensional latent variable. An IIO allows for the arrangement of items from easy to difficult, better interpretation of test scores, and comparison of response patterns. For tests with clustered items, the usefulness of the IIO definition and existing methods for investigation is limited. In this presentation we discuss a) alternative ordering structures in the form of a weak and strong invariant cluster ordering, b) a procedure for investigating the ordering structure of a clustered item set is proposed to guide analysis of test data, and c) the results of this procedure applied to test data from the Norwegian version of the Test for Reception of Grammar. Suggestions for practice, further methodological developments, and future research are discussed.

Integrating observations and test scores: A fixed-random item model

Tuesday, 16th July - 11:30: Developments in IRT (RB 210) - Oral

Prof. Mark Wilson (University of California, Berkeley), Dr. Perman Gochyyev (University of California, Berkeley)

As technology enables the advancement of measurement practices more and more into field observations, the application of psychometric models moves beyond the typical setting of fixed sets of items to contexts where the “items” themselves may vary from situation to situation. An example from the education domain is where teacher in-class observations (i.e., at the “micro” level of grain-size) are to be combined with more typical standardized test outcomes to provide better student ability estimates (i.e., at the “meso” and “macro” levels of grain-size). To this end, we have been developing and applying a more comprehensive characterization of fixed-random item models.

In this paper we analyze micro-macro data where a Construct Map is utilized to provide a criterion-referenced interpretation for pretest and posttest results using a fixed effects model. Then, that interpretive framework is extended to model teacher “on-the-fly” observations (based on the same construct map framework) collected during the school year using a fixed-random item model.

The applications of this approach include: (a) increasing the interpretability of in-class observations using the interpretive framework; (b) augmenting the reliability of observations by combining the observations data with test results; and (c) tracking student progress during the school year.

Issues of estimation, constraint-setting, data sparsity, common scale, and interpretation across different modes of assessment (i.e., macro (standardized test) versus micro (observations)) are discussed. Suggestions are made regarding next steps in terms of both the technology of measurement across different grain-sizes and the development/adaptation of appropriate psychometric models.

A latent variable model for multiple resubmission item responses

Tuesday, 16th July - 11:45: Developments in IRT (RB 210) - Oral

Xiang Liu (Educational Testing Service), Hongwen Guo (Educational Testing Service), Mo Zhang (Educational Testing Service), CHEN Li (Educational Testing Service), Amy Ko (University of Washington), Min Li (University of Washington)

In a learning environment, students are often allowed to respond to the same set of items multiple times, with feedback provided after each submission. In addition to measuring students' proficiency levels (i.e., their ability to correctly answer items on the first try), there is typically interest in assessing their improvement (or learning rate) between multiple submissions. In this presentation, we introduce a latent variable model that addresses this issue. Using the dataset presented, which is collected after an introductory computer science course where students learn Python programming, each student responds to items that require them to write Python programs to complete certain tasks. Their responses are then submitted and evaluated against test cases. After each submission, students receive feedback indicating which test cases they passed and which they failed. An unlimited number of attempts are allowed. We introduce a latent variable model that allows instructors or researchers to profile students in a two-dimensional latent trait space of Python programming proficiency levels and learning rates. Specifically, we will discuss model formulations, parameter estimation, and provide a real data analysis example where the model is fitted to the CS education dataset to demonstrate its utility.

New Developments in Dynamic Modeling

Tuesday, 16th July - 10:45: Symposium: New Developments in Dynamic Modeling (RB 211) - Symposium Overview

Ivan Jacob Pesigan (The Pennsylvania State University)

In recent years, dynamic modeling has witnessed significant progress, yet several critical issues remain unresolved. This symposium aims to shed light on these challenges and foster innovative solutions. Researchers often have to grapple with the challenges associated with analyzing count and discrete-valued time series, especially in fields like psychology and education. Crawford et al. delve into analytic standard errors for stationary vector discrete-valued time series models, accommodating various distributions and flexible autocorrelations. Oh et al. bridge the gap by extending fit indices from structural equation modeling to state space models, considering observed and latent variables, lagged observations, missingness, and different degrees of model complexity. Xiong et al. address the scarcity of discussions on reliability within continuous time (CT) models, extending reliability indices from discrete time (DT) models. Lastly, Pesigan et al. contribute to continuous-time mediation models, facilitating computation of total, direct, and indirect effects along with standard errors and confidence intervals, enhancing our understanding of mediated changes, over time.

Analytic Standard Errors for Latent Gaussian Discrete-Valued Multivariate Time Series

Tuesday, 16th July - 10:45: Symposium: New Developments in Dynamic Modeling (RB 211) - Symposia

Mr. Christopher Crawford (The Pennsylvania State University), Dr. Vladas Pipiras (University of North Carolina - Chapel Hill), Dr. Marie-Christine Düker (Friedrich-Alexander University), Dr. Zachary Fisher (The Pennsylvania State University)

Unlike their continuous-valued counterpart, there are no universally preferred methodologies for modeling count and discrete-valued time series. This is especially problematic in fields like psychology and education where repeated measures data often take the form of count, dichotomous, and ordered categorical variables. To address the need for flexible methodology to analyze discrete-valued time series data, a multivariate model defined via deterministic functions of a latent stationary vector Gaussian series has been proposed. This model has several promising features including the ability to accommodate a wide variety of marginal distributions within the same model (including overdispersed and zero-inflated distributions) while also allowing for the most flexible autocorrelations possible. In this presentation we extend the work on this model by developing analytic standard errors to support inference on the latent dynamics and parameters associated with the marginal distributions. Properties of these analytic standard errors are examined using both empirical and simulated data.

Extending Fit Indices from Structural Equation Models to State Space Models

Tuesday, 16th July - 10:45: Symposium: New Developments in Dynamic Modeling (RB 211) - Symposia

Mr. Hyungeun Oh (The Pennsylvania State University), Dr. Michael Hunter (The Pennsylvania State University), Dr. Sy-Miin Chow (The Pennsylvania State University)

More and more researchers are becoming interested in intensive longitudinal data and the possibilities for modeling within-person processes as they evolve over time. A longstanding method for analyzing these kinds of data from the time series, econometrics, and engineering literature is the discrete-time state space model, which can be thought of as an extension of conventional structural equation modeling to recursively handle large amounts of repeated measurements while allowing for both observed and latent variables. Variations of the state space model have yielded fruitful applications in the social and behavioral sciences.

However, conventional measures of model fit that exist for structural equation models do not exist for the state space model. The purpose of the present work is to extend structural equation modeling fit indices to state space models. Through analytical work and a small simulation study, we evaluate the proposed method of obtaining fit indices for state space models involving: (1) strictly observed variables versus latent and observed variables, (2) various lags between observations, (3) person-and time-specific missingness, and (4) different degrees of model complexity and sample size configurations. We conclude by discussing implications for state space model fit assessment and model selection.

Reliability in Multilevel Continuous Time Modeling

Tuesday, 16th July - 10:45: Symposium: New Developments in Dynamic Modeling (RB 211) - Symposia

Xiaoyue Xiong (The Pennsylvania State University), Dr. Michael Hunter (The Pennsylvania State University), Dr. Zachary Fisher (The Pennsylvania State University), Dr. Sy-Miin Chow (The Pennsylvania State University)

Continuous time (CT) models have become increasingly prevalent in recent years. In contrast to traditional discrete time (DT) models, CT models can readily handle observations measured at unequal intervals and offer the flexibility to explain dynamic and cross-process coupling effects irrespective of the specific lengths of time intervals. Despite their growing popularity, discussion of issues of reliability within the CT framework is sparse. Reliability, the consistency of measurement over repeated tests, is crucial for ensuring similar results across replications. In multilevel longitudinal data, questions of reliability encompass both within-person (consistency over time for the same individual) and between-person (consistency across different individuals) considerations within the DT modeling framework (e.g., Nezlek, 2017; Schuurman & Hamaker, 2019). This study proposes and evaluates extensions of the reliability indices developed for the DT framework to the CT framework. We explore the properties of these reliability indices as effect size measures in power analysis for multilevel irregularly spaced longitudinal data in a Monte Carlo simulation study. Our research highlights ways to conduct reliability analysis in CT modeling, which is vital for quantifying measuring consistency in irregularly spaced, multilevel longitudinal data, and helps hasten our understanding of important design considerations in future intensive longitudinal studies.

Standard Errors and Confidence Intervals for Total, Direct, and Indirect Effects in Continuous-Time Models

Tuesday, 16th July - 10:45: Symposium: New Developments in Dynamic Modeling (RB 211) - Symposia

Ivan Jacob Pesigan (The Pennsylvania State University), Michael Russell (The Pennsylvania State University), Dr. Sy-Miin Chow (The Pennsylvania State University)

Mediation Models Mediation modeling using intensive longitudinal data is an exciting field that captures the interrelations in dynamic changes, such as mediated changes, over time. Even though discrete-time vector autoregressive (DT-VAR) approaches are commonly used to estimate indirect effects in intensive longitudinal data (ILD), they have known limitations due to the dependency of inferential results on the time intervals between successive occasions and the assumption of regular spacing between measurements. To address these issues, continuous-time vector autoregressive (CT-VAR) models have been proposed as an alternative. Previous work in the area (e.g., Deboeck & Preacher, 2015 and Ryan & Hamaker, 2021) has shown how the total, direct, and indirect effects, for a range of time-intervals values, can be calculated using parameters estimated from CT-VAR models for causal inferential purposes. However, methods for calculating the uncertainty around the total, direct, and indirect effects in continuous-time mediation have yet to be explored. Drawing from the mediation model literature, we present and compare results from using the delta and Monte Carlo methods to calculate standard errors and confidence intervals for the total, direct, and indirect effects in continuous-time mediation for inferential purposes. Options to automate these inferential procedures and facilitate interpretations are available in http://github.com/ijapesigan/cTMed.

Hierarchical relation among minimum rank FA, PCA, and prevalent FA

Tuesday, 16th July - 10:45: Topics in Factor Analysis (RB 212) - Oral

Prof. KOHEI ADACHI (Osaka University)

A prevalent procedure of factor analysis (PrevFA) follows from the FA model in which a multivariate observation is decomposed into two parts: common factor (CF) and unique factor (UF) (e.g., Mulaik, 2010). In a more comprehensive FA (CompFA) model (Adachi, 2022), the observation is decomposed into three parts: CF, UF, and error. The CompFA model leads to the error covariance matrix that is equal to the sample covariance matrix minus the sum of the loading matrix post-multiplied by its transpose and the diagonal matrix including UF variances.

I discuss how ten Berge and Kiers' (1991) minimum rank FA (MRFA), which is an alternative to PrevFA, can be reformulated as an estimation procedure for the CompFA model. In more detail, I show that MRFA can be regarded as minimizing the trace of the error covariance matrix subject to its being proper, i.e., nonnegative definite. Then, I discuss how principal component analysis (PCA) can be reformulated as a procedure for the constrained CompFA (C-CompFA) model with the UF variances restricted to zeros. Here, PC scores are treated as latent variables. Finally, the model for PrevFA is shown to be the constrained version of the C-CompFA model with the error covariance matrix restricted to a diagonal one. Thus, we can have the hierarchical relation MRFA > PCA > PrevFA, where the model underlying the procedure after > is a constrained version of the model underlying the one before >.

The partially oblique rotation for exploratory factor analysis

Tuesday, 16th July - 11:00: Topics in Factor Analysis (RB 212) - Oral

Prof. Marcos Jiménez (Universidad Autónoma de Madrid), Dr. Francisco J. Abad (Universidad Autónoma de Madrid), Dr. Eduardo García-Garzón (Universidad Camilo José Cela), Dr. Luis Eduardo Garrido (Pontificia Universidad Católica Madre y Maestra), Dr. Vithor Franco (Universidade São Francisco)

Over the past two decades, orthogonal and oblique rotations has been the primary methods for interpreting factor solutions in exploratory factor analysis. However, there are situations where researchers have prior knowledge that certain factor correlations should be zero while others could be non-zero, like in bi-factor models featuring multiple general factors and multitrait-multimethod models. In these cases, researchers are forced to switch to confirmatory factor analysis, which can offer poor fit for large and complex factor structures. Thereby, we have introduced a novel rotation approach known as partially oblique rotation, in which both oblique and orthogonal factors co-exist. This rotation allows for estimating exploratory versions of both bi-factor models with multiple general factors and multitrait-multimethod models. Additionally, we derived the rotation constraints that are required for computing standard errors. We illustrated the results of this rotation method through examples drawn from personality and education research. Finally, we provided an R package that facilitates the implementation of the new partially oblique rotation.

Overcoming Heywood cases in factor analysis: a maximum penalized likelihood approach

Tuesday, 16th July - 11:15: Topics in Factor Analysis (RB 212) - Oral

Mr. Philipp Sterzinger (University of Warwick), Prof. Ioannis Kosmidis (University of Warwick), Prof. Irini Moustaki (The London School of Economics and Political Science)

Exploratory Factor Analysis (EFA) is a widely used statistical method in psychometrics for uncovering latent structures from observed data. A persistent issue in EFA models is the emergence of Heywood cases, where estimates fall at the boundary of the parameter space. These cases challenge the validity and interpretability of the analysis.

We introduce a maximum penalized likelihood approach to parameter estimation in EFA that addresses this issue. Specifically, we penalize the log-likelihood additively by appropriately scaled penalty functions on the factor loadings and unique variances. The penalties are designed to diverge to minus infinity for parameters at the boundary, effectively avoiding Heywood cases. Appropriate scaling ensures that the bias of the estimates away from the MLE, which is introduced from the penalties, is asymptotically negligible.

Thus, our framework guarantees the existence of estimates in the interior of the parameter space, effectively avoiding Heywood cases. Importantly, it maintains key statistical properties: consistency, asymptotic normality, and rotational invariance under orthogonal rotations of the factor loadings. Regularized estimation to prevent Heywood cases in EFA is not a novel concept; our framework is, however, the first to provide formal guarantees of its effectiveness and performance.

We conduct a series of simulations to evaluate the framework. These simulations focused on three aspects: the accuracy of the estimators in finite samples, the concordance of their finite sample performance with our asymptotic theory and the model selection capability of the framework. The results indicate improved handling of Heywood cases and enhanced reliability in factor analysis outcomes.

Does choice of nonnormal data generation algorithm affect simulation results?

Tuesday, 16th July - 11:30: Topics in Factor Analysis (RB 212) - Oral

Prof. Amanda Fairchild (University of South Carolina), Mr. Yunhang Yin (University of South Carolina), Prof. Oscar Astivia (University of Washington), Prof. Dexin Shi (University of South Carolina), Prof. Amanda Baraldi (Oklahoma State University)

Rationale:

Discussions around replicability have been limited largely to empirical studies. Few researchers have advocated replicating methodological work. This is a critical gap, given a single Monte Carlo simulation can change the conclusions of thousands of empirical studies as new guidelines and best practices are developed.

This meta-scientific work addresses that gap by replicating one widely cited methodological study (Curran et al., 1996) and evaluating generalizability of results across different nonnormal data generation algorithms. Our work speaks to the broader notion that simulation study recommendations are sensitive to design choices.

Method:

We generated multivariate nonnormal data for a correctly specified, three-factor CFA model. All analyses were conducted using R 4.3.1. Nonnormal data generation algorithms examined were: VM, Headrick, IG, PLSIM, GC Normal, GC Chi, and NORTA. Estimators examined were normal theory ML and robust MLM. Level of nonnormality (moderate vs. severe) and sample size (N=200, 500, 1000) were also manipulated. Outcomes examined were relative bias of the model chi-square and empirical rejection rates.

Results:

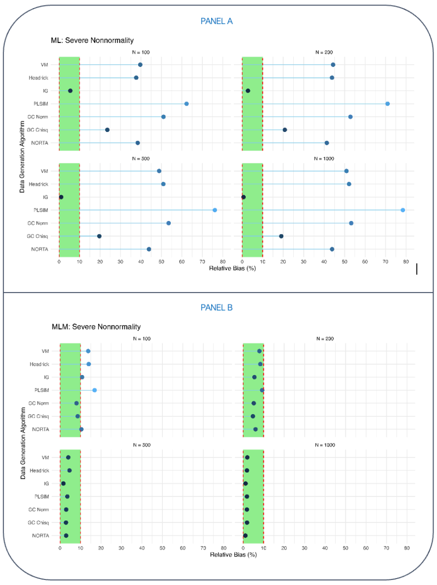
Replication results were generally consistent with Curran et al., but several differences in magnitude were discerned. Generalizability results were more mixed. Only two results observed under the original data generation algorithm held completely across other algorithms examined. Several other results generally held across most of the comparator algorithms.

Implications:

Findings suggest that existing methodological recommendations may not be universally valid in circumstances where more than one data generation algorithm is available to simulate a given data characteristic. We offer recommendations for both applied and methodological researchers.

Figure 1.

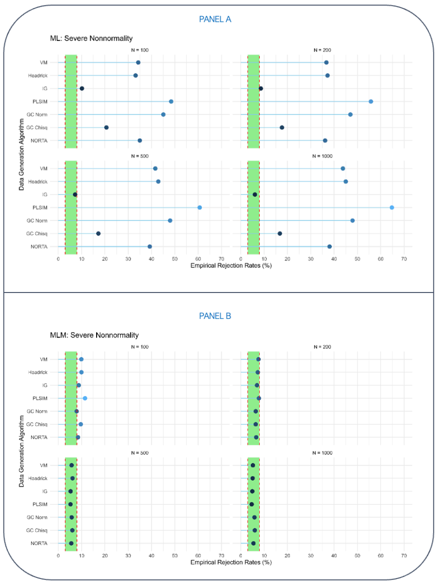
RB of model χ^2 across data generation algorithm and N under severe nonnormality



Fairchild et al. figure 1 imps 2024.png

Figure 2.

Empirical rejection rate across data generation algorithm and N under severe nonnormality



Fairchild et al. figure 2 imps 2024.png

Continuous-time latent variable modeling of measurement scales in health studies: Applications to the Multiple-System Atrophy progression

Tuesday, 16th July - 13:30: Invited Talk (Vencovského aula) - Invited Talk

Dr. Cécile Proust-Lima (University of Bordeaux and Inserm)

The study of neurodegenerative diseases, such as the Multiple System Atrophy (MSA: a rare alpha-synucleinopathy with poor prognosis) using data from epidemiological cohorts presents various statistical challenges:

- There is a large heterogeneity in the typical profiles of progression;
- Multiple dimensions/processes that evolve over time are involved;
- Most dimensions are measured by scales made of continuous or ordinal items (e.g., cognitive functions, motor function, quality of life).
- Repeated measurements are collected at highly variable times across participants, with missing data.
- Events (e.g., diagnosis, death, dropout) truncate the observation process.

Through the analysis of a large cohort of MSA patients, I describe several methodologies based on latent variables (random effects, latent processes, latent classes) to model disease progression in continuous time while addressing these challenges. Specifically, I introduce a continuous-time item response model (for continuous and/or ordinal items) to handle the irregular and individual-specific timings of observation in cohorts (1,2) and its extension to the joint modeling of a clinical event (3). Then, I describe an approach to study the progression of health-related Quality-of-Life over time along the clinical progression (4). Finally, I investigate the phenotypic heterogeneity via a latent class model for multiple longitudinal processes and time-to-death (5). These approaches are available in open-source R software, notably in the lcmm R package (6).

Confirmatory mixture models for investigating contextual correlates of response behavior in ecological momentary assessments

Tuesday, 16th July - 13:30: Invited Talk (RB 101) - Invited Talk

Dr. Esther Ulitzsch (Centre for Educational Measurement (CEMO), University of Oslo)

In ecological momentary assessment (EMA) studies, respondents report on their current behaviors and experiences on several occasions throughout the day, for multiple days. This in-depth data collection offers a unique window into human behavior and experiences. The validity of conclusions drawn from EMA data, however, rests on the assumption that respondents interact with the administered measures in the same way on all measurement occasions. In this talk, I focus on inattentive responding as a prominent example of deviating interactions. When inattentive, respondents provide their responses without investing effort into carefully evaluating the items. As a result, EMA data may be contaminated with responses that do not reflect what researchers want to measure. I illustrate how translating subject-matter theory on respondent behavior into confirmatory mixture models facilitates studying the occurrence of deviating interactions as well as their contextual correlates (e.g., time of day), discussing both item responses and process data as data sources that can be used for model formulation. I end with a brief note on other types of heterogeneous respondent interactions that could be captured and investigated with appropriately formulated confirmatory mixture models.

Decoding intelligence: investigating the structure of intelligence

Tuesday, 16th July - 14:30: Psychometric Applications to Psychology (RB 101) - Oral

Ms. Laura Maria Fetz (University of Amsterdam)

The present study explores the structure of intelligence across different countries by utilizing meta-analytical structural equation modeling (MASEM) and meta-analytical Gaussian network analysis (MAGNA). The data set is unique in a number of aspects. First, used intelligence test is designed to minimize potential cultural bias by simplifying language and reducing the reliance on academic knowledge as much as possible. Second, the data set is big (for the field of intelligence) and includes data from roughly 150.000 participants from over 200 countries. Third, the test is computer adaptive, which results in a high number of missing. Our research aims are threefold: (1) By comparing the mainstream higher order factor model of intelligence with a psychometric network of intelligence, we aim to replicate or challenge recent findings which suggest that network model outperform the mainstream factor model. (2) We aim to understand whether the applied meta-analytical techniques are suitable and feasible. (3) By employing a Monte Carlo Simulation, we address possible bias in parameter estimation. This research therefore has the potential to significantly contribute to the field of both psychometrics and cross-cultural social science.

Keywords: Intelligence, hierarchical factor modeling, psychometric network modeling, Meta-Analytical Structural Equation Modelling (MASEM), Meta-analytical Gaussian network analysis (MAGNA), Monte Carlo simulation

An investigation the context effect in the multidimensional forced-choice personality measurement

Tuesday, 16th July - 14:45: Psychometric Applications to Psychology (RB 101) - Oral

Dr. Kyosuke Bunji (Kobe University), Dr. Kensuke Okada (The University of Tokyo)

A fundamental assumption in item response theory (IRT) models for multidimensional forced-choice (MFC) personality measurement is that the utility of each statement (choice option) remains independent and invariant, irrespective of factors such as the context effect. However, investigations into the confounding impact of the context effect on MAFC personality measurements remain scarce. This study examined the impact of the context effect on the utility and estimation of the parameters of each statement in the MFC personality measurement.
We postulated the same statement in different item blocks and estimated parameters using the Thurstonian IRT model, assuming either the absence (item parameters of the same statement are identical even when appearing in different blocks) or presence (item parameters of the same statement vary in different blocks) of context effects. We checked the estimates from various perspectives and obtained the following results:
1. The correlation of trait scores is almost one, regardless of the existence of the context effect.
2. The difference in an information criterion (WAIC) was small (within about 10% of the standard error).
3. The 95% credible interval of “difference” parameters included zero.
4. In terms of the Bayes factor, the model without the context effect was the best.
5. The overlap rate of the posterior distributions of the same statement was about 80%.
The study’s results suggest that the context effect can be considered small enough to be practically negligible.

Dynamic and bidirectional associations between physical-activity, smartphone addiction and state self-esteem

Tuesday, 16th July - 15:00: Psychometric Applications to Psychology (RB 101) - Oral

Yang Cui (Faculty of Psychology, Beijing Normal University), Manlu Zhang (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China), Chengwei Zhu (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China), Prof. Cai Zhang (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China), Prof. Fumei Chen (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China), Li Ke (Faculty of Psychology, Beijing Normal University), Prof. Yun Wang (Faculty of Psychology, Beijing Normal University), Prof. Danhui Zhang (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China)

Smartphones play a crucial role in people's daily life with lots of benefits. The growing popularity of smartphone has led to smartphone overuse and addiction among adolescents. Previous studies suggested that problematic smartphone use or addiction associated with physical and mental health problems, including alcohol use disorder, depression, anxiety, and low self-esteem. Physical activity is associated with many positive outcomes across development. A significant negative association has been found between physical activity and smartphone addiction. Moreover, series of studies found that physical activities play a powerful role in the development of self-esteem. However, the bidirectional and longitudinal dynamics of how physical activity, smartphone addiction and state self-esteem influence each other based on intensive longitudinal data remain unclear.

To address the gap in the literature, college students (N=239, Mage = 21.18 years, 79% female) completed questionnaires twice a week over a span of 12 weeks reporting their smartphone use (Mobile Phone Problematic Use Scale, MPPUS), state self-esteem (Six-Item State Self-Esteem Scale, SSES-6) and the frequency of physical activity (Godin Leisure Time Exercise Questionnaire, GLTEQ). Based on the dynamic structural equation modeling (DSEM), we examined the within-person autoregressive and cross-lagged associations among state self-esteem, addiction, and physical activities.

The results show that previous physical activities can predict subsequent state self-esteem positively among college students whereas addiction was negatively associated. Smartphone addiction was negatively associated with state self-esteem only on days without physical exercise. Findings suggest that physical activity buffers the lower state self-esteem associated with addiction.

Predicting onset of child positive and negative affect and behavior

Tuesday, 16th July - 15:15: Psychometric Applications to Psychology (RB 101) - Oral

Dr. Jenn-Yun Tein (Arizona State University), Dr. Cheuk-Hei Bosco Cheng (Arizona State University), Dr. Sierra Clifford (Arizona State University), Ms. Sophia Lamp (Arizona State University), Dr. Daniel Shaw (University of Pittsburgh), Dr. Melvin Wilson (University of Virginia), Dr. Kathryn Lemery-Chalfant (Arizona State University)

Dynamic parent-child interactions in early childhood are central processes contributing to child development and mental health. Direct observations with real-time coding of family interactions systematically capture both maladaptive and healthy dynamic behavioral patterns, as well as their antecedents and consequences. Understanding early childhood parent-child interaction can uncover nuances and subtleties of family dynamics and their relations to the future course of youth substance use and dependence (SUD) and mental health problems. This project used second-by-second coded parent-child interaction observational data from a randomized controlled trial of the family-centered Family Check-Up (FCU) intervention (N = 731) to study how family and child factors predict children's positive and negative affect and behavior at ages 2-4. Specifically, we applied multilevel survival analysis to examine patterns of onset of child negative and positive behavior. In addition, we examined how the behavior pattern related to time-invariant predictors (child gender, family cumulative risks, parental depression) and time-variant predictors (parent positive and negative behaviors prior to each onset of child behavior). A preliminary analysis showed that child word understanding and baseline internalizing problems, as well as parent age and specific behaviors during the interaction, were related to child positive and negative response patterns. Using multilevel survival analysis to understand detailed parent-child engagement processes and their relations to child and family characteristics can inform the development of tailored and effective intervention strategies.

Developing a framework for testing dyadic associations using pseudodyads and permutations

Tuesday, 16th July - 15:30: Psychometric Applications to Psychology (RB 101) - Oral

Chiara Carlier (KU Leuven), Liesse Frérart (KU Leuven), Prof. Peter Kuppens (KU Leuven), Prof. Eva Ceulemans (KU Leuven)

Social interactions are an important aspect of our lives. In many of these interactions, we form a unit of two, a dyad. Psychological research has recently gained more interest in dyadic relationships, investigating among others the extent to which minority individuals show similar values as cultural majority peers or whether romantic partners synchronize their emotions and behaviors. Assessing these associations between psychological constructs such as emotions and values requires appropriate statistical tools that take into account the correct baseline. Such a baseline is expected to be zero when using correlation measures as association index, but is less well defined in distance measures. In addition, random persons always share some level of similarity due to similar backgrounds, circumstances, or surroundings, leading to a possibly different baseline correlation than zero. To infer significance of dyadic associations in comparison to the correct baseline, R.J. Corsini developed and D.A. Kenny further discussed the method of forming pseudo couples, by pairing each dyad member in a given sample with an unrelated individual from another dyad and recomputing the association between those randomly paired partners. In this talk we will elaborate on the development of a framework, including Shiny app, for enabling the testing of dyadic associations with the help of pseudo couples. Additionally, we will extend this permutation method that is tailored to cross-sectional data to intensive longitudinal data.

Ordinal latent space item response models for Likert scale assessments.

Tuesday, 16th July - 14:30: Topics in Categorical Data Analysis (NB A) - Oral

Ludovica De Carolis (University of Milano-Bicocca and University of California Los Angeles), Dr. Inhan Kang (Yonsei University), Minjeong Jeon (University of California, Los Angeles)

A latent space item response model (LSIRM, Jeon, Jin, Schweinberger, & Bauch, 2021) has been proposed to model and explore conditional dependencies or item-by-person interactions in binary item response data. We propose an extension of the LSIRM tailored for ordinal Likert scale data from psychological assessment. The ordinal LSIRM captures unobserved item-by-person interactions in Likert-scale assessment data in the form of distances between persons and items in an interaction map, where the item-by-person distances can be translated into the person's likelihood of endorsing the items given their overall trait levels. We will show how the interaction map can be utilized to derive personalized diagnostic feedback for individuals, e.g., in terms of their likelihood of showing the symptoms in a clinical assessment. We will validate the proposed approach and further demonstrate its utility over traditional ordinal item response models as well as the original LSIRM for binary response data via simulations and empirical studies.

Implications of alternative parameterizations in structural equation models for longitudinal categorical variables

Tuesday, 16th July - 14:45: Topics in Categorical Data Analysis (NB A) - Oral

Prof. Silvia Bianconcini (University of Bologna), Kenneth A. Bollen (University of North Carolina at Chapel Hill)

When analyzing scaling conditions in latent variable Structural Equation Models (SEMs) with continuous observed variables, analysts scaling a latent variable typically set the factor loading of one indicator to one and either set its intercept to zero or the mean of its latent variable to zero. When binary and ordinal observed variables are part of SEMs, the identification and scaling choices are more varied and multifaceted. This is further complicated by longitudinal data. In SEM software, such as lavaan and Mplus, fixing the underlying variables' variances or the error variances to one are two primary scaling conventions. As we show in this paper, the choice between these constraints can have substantial consequences in longitudinal analysis, leading to differences in model fit and degrees of freedom and influencing assumptions about the dynamic process and error structure. We explore alternative parameterizations and conditions of model equivalence with categorical repeated measures. More specifically, we offer insight into the specifications of the autoregressive latent trajectory model and its special cases, the linear growth curve and first-order autoregressive models, for longitudinal categorical indicators with implications for an even wider variety of longitudinal models.

Similarity measures between two sets of binary vectors

Tuesday, 16th July - 15:00: Topics in Categorical Data Analysis (NB A) - Oral

Prof. Sophie Vanbelle (University of Maastricht), Ana Perišić (University of Split)

Assessing similarity in the context of binary classification is common in psychology (e.g., evaluating the presence/absence of positive affect). Several statistical measures were developed to measure the similarity between two binary vectors (e.g., between one human and one algorithmic classification). For instance, the simple matching coefficient quantifies the similarity between positive and negative matches while the Jaccard index focuses only on positive or negative matches. In the example, considering one particular human and one algorithm limits the ecological validity. We may be interested in comparing humans and algorithms in general, two groups of raters or two sets of algorithms. We therefore propose to extend the simple matching coefficient and the Jaccard index to the case of two sets of binary vectors (e.g., to measure the similarity between the classification made by a group humans versus a set of algorithms). Note that our method can also be applied to other similarity coefficients. We derived the large sample variances of the new coefficients and present real-world applications.

Fitting and testing log-linear item response models with known support

Tuesday, 16th July - 15:15: Topics in Categorical Data Analysis (NB A) - Oral

Dr. David Hessen (Utrecht University)

In this presentation, the support of the joint probability distribution of item scores is treated as unknown. From a general total population model with unknown support, a general subpopulation model with its support equal to the set of all observed score patterns is derived. In maximum likelihood estimation of the parameters of any such subpopulation model, the evaluation of the log-likelihood function only requires the summation over a number of terms equal to at most the sample size. The parameters of a hypothesized total population model turn out to be consistently and asymptotically efficiently estimated by the values that maximize the log-likelihood function of the corresponding subpopulation model. New likelihood ratio goodness of fit tests are presented as alternatives to the Pearson chi-square goodness of fit test and the likelihood ratio test against the saturated model. The results of a simulation study on the asymptotic bias and efficiency of maximum likelihood estimators and the asymptotic performance of the goodness of fit tests are presented.

Beyond ignorability: Advancing with Partial Identifiability in Psychometrics

Tuesday, 16th July - 14:30: Symposium: Beyond ignorability: Advancing with Partial Identifiability in Psychometrics (NB B) - Symposium Overview

Dr. Eduardo Alarcón-Bustamante (Pontificia Universidad Católica de Chile)

In empirical research, the challenge of parameter identification arises when facing with missing data, an issue often encountered within the field of psychometrics. Traditionally, this problem is tackled by using the ignorability assumption, which presupposes that both the observed and non-observed distribution of the variable of interest are the same. This symposium aims to show an alternative pathway for facing parameter identification dilemmas by using weaker and less restrictive assumptions.

We will spotlight the methodology of partial identification, an approach that avoids reliance on strong distributional assumptions, so that inferences are drawn based on the context of the problem. This technique acknowledges the uncertainty inherent due to the missing data without overreaching the limits of the available information, providing more robust and credible results.

The topics to be included in the symposium will be the following:

- Estimation of the marginal effect under partial observability of the outcome in a regression setting.
- Estimation of regression coefficients with missing outcomes, where the goal is to assess the degree to which one variable can predict an outcome of interest.
- Prediction of Grade Point Average (GPA) in the university selection context, a common challenge in educational assessment.
- Principles of Causal Inference, aiming to clarify how causal relations can be inferred from the data.

The speakers will be Ernesto San Martín, Jorge González, Inés M. Varas, and Eduardo Alarcón-Bustamante, all from the Pontificia Universidad Católica de Chile.

A Partial Identifiability Approach for Analyzing Regression Functions with Missing Data in the Outcome

Tuesday, 16th July - 14:30: Symposium: Beyond ignorability: Advancing with Partial Identifiability in Psychometrics (NB B) - Symposia

Dr. Jorge Gonzalez (Pontificia Universidad Católica de Chile), Dr. Eduardo Alarcón-Bustamante (Pontificia Universidad Católica de Chile), Dr. David Torres Irribarra (Pontificia Universidad Católica de Chile), Dr. Ernesto San Martín (Pontificia Universidad Católica de Chile)

In this talk, we present an approach based on the theory of partial identifiability that considers a variety of assumptions to learn about a regression function with missing data in the outcome variable. We use real data on a Chilean college admissions process to illustrate how results can vary with the considered assumptions about the selection process and compare these results with those obtained using more traditional approaches that are based on ignorability and strong distributional assumptions.

On the impact of missing outcomes in predictive capacity of selection tests studies

Tuesday, 16th July - 14:30: Symposium: Beyond ignorability: Advancing with Partial Identifiability in Psychometrics (NB B) - Symposia

Dr. Inés Varas (Pontificia Universidad Católica de Chile), Dr. Eduardo Alarcón-Bustamante (Pontificia Universidad Católica de Chile), Dr. Ernesto San Martín (Pontificia Universidad Católica de Chile)

Linear regression models are commonly used for measuring the impact of covariates over an outcome of interest, which is typically measured through the regression coefficients of the model. For instance, in predictive validity studies the regression coefficients are used for measuring the strength of the relationship between test scores and any outcome of interest (e.g., the performance of students in a university). In the selection context, the test scores are observed for all the applicants however the performance is observed only in selected ones. There is an inherent missing data problem which can seriously affect the interpretation of the regression coefficients. This talk examines the effect of missing outcome data on the relationship between selection factors and student performance in a Chilean university admission process using a partial identification approach.

Assessing the marginal effect under partial observability in a selection context

Tuesday, 16th July - 14:30: Symposium: Beyond ignorability: Advancing with Partial Identifiability in Psychometrics (NB B) - Symposia

Dr. Eduardo Alarcón-Bustamante (Pontificia Universidad Católica de Chile), Dr. Jorge Gonzalez (Pontificia Universidad Católica de Chile), Dr. Ernesto San Martín (Pontificia Universidad Católica de Chile), Dr. David Torres Irribarra (Pontificia Universidad Católica de Chile)

In a university selection process, the test scores are observed for all the applicants. However, the performance in the university can be observed only for those selected ones. This partial observability poses a challenge in assessing the effect of the scores over the performance because assumptions regarding the non-observed effects are necessary. The question is what type of assumptions can be made? In this talk, we explore how to learn about the marginal effect by considering several assumptions about the non-observed effects. We compare conclusions based on distributional assumptions with those based on assumptions related to the selection context. Also, we explore an interpretation of the marginal effect as a function of the scores, providing insights for assessing the predictive capacity of selection tests.

Novel advances in psychometric network modeling

Tuesday, 16th July - 14:30: Symposium: Novel advances in psychometric network modeling (NB C) - Symposium Overview

Mr. Xinkai Du (University of Oslo)

The field of network psychometrics emerged from the network perspective on psychology, which, instead of treating observable variables as merely the indicators of latent common causes, conceptualizes relations among observables (e.g., symptoms, cognitions, and behaviors) as the key to the emergence of psychological phenomena. Upon their advent in psychological studies, psychological networks have been adopted as practical tools, enriching the array of existing instruments used in psychometrics. Network models are powerful tools to visualize and investigate the relations among variables / items, often uniquely identified, and are well-suited for the purpose of cluster detection when no theories on the organization of variables are available. This symposium introduces several latest advancements in the field of network psychometrics: Tuo Liu will discuss the possibility of extending the current Item Response Theory (IRT) based computer adaptive testing (CAT) with Ising Models. Xinkai Du will explore the possibility to progress network psychometrics beyond its exploratory stage, discussing the usefulness and pitfalls of Structural Equation Model (SEM) fit measures in confirmatory network psychometrics. In the next two talks, Ria Hoekstra and Björn Siepe will introduce their novel methodology to determine heterogeneity in idiographic networks, from the perspective of invariance testing and Bayesian statistics respectively. Tatiana Kvetnaya will discuss the efficacy of graphical Gaussian mixture modeling for exploratory subgroup identification in psychological networks, focusing on the impact of ordinal and skewed data scenarios. Finally, Kai Nehler will discuss the handling of missing data in network estimation.

From exploratory to confirmatory network psychometrics: Fit indices and cutoff values

Tuesday, 16th July - 14:30: Symposium: Novel advances in psychometric network modeling (NB C) - Symposia

Mr. Xinkai Du (University of Oslo), Mr. René Freichel (University of Amsterdam), Ms. Nora Skjerdingstad (University of Oslo), Mr. Omid Ebrahimi (University of Oxford), Ms. Ria Hoekstra (University of Amsterdam), Dr. Sacha Epskamp (National University of Singapore)

Psychometric networks allow for the investigation of inter-variable relations through uniquely identified multivariate models, without the assumption of existing latent variables. Thus, network models are well-suited for phenomena detection, and most empirical network studies have been exploratory in nature. Yet, for the close connections between (Gaussian) network modeling and Structural Equation Modeling (SEM), techniques in SEM literature are readily transferable to network modeling as well. Therefore, despite the largely exploratory nature of network analysis research so far, confirmatory network psychometrics has been entirely feasible. However, no study to date has evaluated how confirmatory network analysis should be implemented, and what criteria should be applied to such tests. In this study, we explored the possibility of utilizing SEM fit indices to detect misfit in confirmatory network analysis, with panel graphical autoregressive model as a representative, for its generalizability to both models used in cross-sectional (Gaussian graphical models) and N = 1 time-series case (graphical autoregressive). We analyzed the sensitivity of fit indices to different forms of model misspecification, varying the number of nodes, sample size, and number of waves. Most fit indices performed well except for PNFI. The conventional cutoffs in SEM literature are largely generalizable to confirmatory network analysis to be a convenient assessment criteria, when dynamical cutoffs are unavailable. We also discussed fit indice's vulnerability to the covariates controlled. The study aims to motivate confirmatory network psychometrics and encourage theory-testing research in network studies, providing a guideline for the use of SEM fit indices in confirmatory network testing.

Integration of Ising Model and Computer Adaptive Testing: A Simulation

Tuesday, 16th July - 14:30: Symposium: Novel advances in psychometric network modeling (NB C) - Symposia

Mr. Tuo Liu (Goethe University Frankfurt), Dr. Aron Fink (Goethe University Frankfurt), Dr. Sacha Epskamp (National University of Singapore)

Currently, computerized adaptive testing (CAT) relies heavily on item response theory (IRT). Network psychometrics provides an alternative, considering the measurement of a stable organization of dynamic, interacting components. An adaptive form of network psychometrics has been developed, including a network-based item selection algorithm and an Ising model-based estimation method. More specifically, this item selection algorithm would administrate the item, which could maximally reduce the entropy to predict the responses of unanswered items given the already answered items. This process would be repeated until all remaining unanswered items can adequately be predicted using an Ising model instead of the IRT model.

Empirical studies on this adaptive form of network psychometrics are limited. This research compares its performance to CAT by examining item selection algorithms and model estimation methods. A simulation based on previous psychometric data employs a factorial design encompassing the item-selection factor (Kullback-Leibler divergence criterion, Fisher information criterion, and Entropy reduction criterion), model estimation factor (IRT and Ising), and the number of administered items. As the dependent variables, the proportion of misclassification error will be compared among all combinations of factors. As an exploratory approach, the results will provide detailed insights into the complementary application potential of the adaptive form of network psychometrics.

Testing for similarity between idiographic networks: The Individual Network Invariance Test (INIT)

Tuesday, 16th July - 14:30: Symposium: Novel advances in psychometric network modeling (NB C) - Symposia

Ms. Ria Hoekstra (University of Amsterdam)

In many applied settings, the task of comparing idiographic network structures has been a challenging endeavor. Previously, researchers resorted to methods such as eyeballing the estimated network structures or deployed techniques that make use of the multilevel structure of the data. Although these methods have their benefits, they are limited by a significant drawback: their inability to rigorously assess whether individual network structures are similar or different. To bridge this gap, we introduce the Individual Network Invariance Test (INIT), an innovative approach designed to facilitate the direct comparison of idiographic network structures. INIT extends well-used comparison practices within Structural Equation Modeling (SEM) to the realm of idiographic network analysis. The performance of INIT using simulations will be highlighted on both saturated (i.e., fully connected) and pruned (i.e., some of the matrix elements have been set to zero) idiographic network structures. In addition, the possibilities of this new technique will be illustrated, highlighting how INIT allows testing not just for (in)equality between idiographic network structures, but also within idiographic network structures.

Bayesian Estimation and Comparison of Idiographic Network Models

Tuesday, 16th July - 14:30: Symposium: Novel advances in psychometric network modeling (NB C) - Symposia

Björn Siepe (Psychological Methods Lab, Department of Psychology, University of Marburg), Mr. Matthias Kloft (Psychological Methods Lab, Department of Psychology, University of Marburg), Prof. Daniel Heck (Psychological Methods Lab, Department of Psychology, University of Marburg)

Idiographic network models are estimated on time-series data of a single individual and allow researchers to investigate person-specific associations between multiple variables over time. The most common approach for fitting graphical vector autoregressive (GVAR) models uses LASSO regularization to estimate a contemporaneous and a temporal network. However, estimation of idiographic networks can be unstable in relatively small data sets typical for psychological research. This bears the risk of misinterpreting differences in estimated networks as spurious heterogeneity between individuals. As a remedy, we evaluate the performance of a Bayesian alternative for fitting GVAR models that allows for regularization of parameters while accounting for estimation uncertainty. We also develop a novel test, implemented in the *tsnet* package in R, which assesses whether differences between estimated networks are reliable based on matrix norms. We first compare Bayesian and LASSO approaches across a range of conditions in a simulation study. Overall, LASSO estimation performs well, while a Bayesian GVAR without edge selection may perform better when the true network is dense. In an additional simulation study, the novel test is conservative and shows good false-positive rates. Finally, we apply Bayesian estimation and testing in an empirical example using daily data on clinical symptoms for 40 individuals. We additionally provide functionality to estimate Bayesian GVAR models in Stan within *tsnet*. Overall, Bayesian GVAR modelling facilitates the assessment of estimation uncertainty which is important for studying inter-individual differences of intra-individual dynamics. In doing so, the novel test serves as a safeguard against premature conclusions of heterogeneity.

Exploratory subgroup detection of psychological networks: Assessing the impact of ordinal and skewed data

Tuesday, 16th July - 14:30: Symposium: Novel advances in psychometric network modeling (NB C) - Symposia

Tatiana Kvetnaya (Goethe University Frankfurt), Kai Jannik Nehler (Goethe University Frankfurt), Martin Schultze (Goethe University Frankfurt)

Exploratory subgroup identification can be a valuable tool for psychological network science, e.g., to identify patient subgroups with distinct symptom constellations in mental disorders. Gaussian mixture modeling (GMM) – a popular method for investigating heterogeneity in multivariate data – offers a promising avenue to achieve this. GGM approaches allow observations to be clustered into subgroups based on their network characteristics, rather than symptom profiles or sum scores.

Recent advancements in graphical GMM approaches were extended to explicitly consider the structure of associations among variables within each cluster (e.g., Fop et al. 2018). By introducing a graph structure search step into the expectation–maximization (EM) algorithm, it allows for not only optimizing parameters but also graph edge sets. However, this approach assumes continuous, normally distributed data, whereas real-world psychological data is often ordinal and/or skewed in nature.

In this presentation, we seek to explore how effectively the structural EM algorithm is able to recover underlying subgroups in data under conditions frequently encountered in psychological data. To this end, we generate cross-sectional data stemming from 3 subgroups with different degrees of network sparsity, echoing findings from previous network analyses of psychological disorders. By varying the cluster proportions, the number of ordinal answer categories, and variable skewness in the simulated datasets, we evaluate the performance of graphical GGM in terms of clustering and structure recovery. Classification goodness, as well as recovery of the true cluster proportions, edge sets, and weight estimates are used as performance indicators.

Challenges in Sample Size Calculation for Psychological Networks with Missing Data

Tuesday, 16th July - 14:30: Symposium: Novel advances in psychometric network modeling (NB C) - Symposia

Kai Jannik Nehler (Goethe University Frankfurt), Martin Schultze (Goethe University Frankfurt)

The initial phase of developing and comparatively evaluating methods for estimating psychological networks in cross-sectional settings primarily focused on fully observed data (e.g., Isvoranu & Epskamp, 2023). In the presence of missing data, analysis software often defaulted to either throwing errors, employing simplistic missing data handling, or utilizing more advanced methods without thorough evaluation. With a growing body of research interest in this field, a number of inconsistencies in missing data handling has surfaced. One of these is the determination of the sample size, which is pivotal within estimation techniques such as the most popular graphical lasso regularization (Friedman et al., 2008) or non-regularized alternatives like neighborhood selection (Williams et al., 2020). In this talk, our focus is on investigating the implications of choosing different approaches to define the sample size within various estimation techniques. We first present currently implemented defaults in standard software packages and then delve into the performance of approaches to define network size via a simulation study, evaluating a variety of settings (e.g., degree of missingness or sample size). Evaluation metrics include the identification of the edge set and examining bias in partial correlations and strength values.

Statistical learning approaches to psychometric challenges

Tuesday, 16th July - 14:30: Symposium: Statistical learning approaches to psychometric challenges (NB D) - Symposium Overview

Dr. Dylan Molenaar (University of Amsterdam)

Due to the rapid recent developments in the fields of artificial intelligence, many interesting new statistical modeling tools have become available. In this symposium, we explore the usefulness of these modeling tools for the field of psychometrics. That is, we focus on various practical cross-sectional and longitudinal psychometric challenges, and propose solutions that combine models and algorithms from the fields of statistical learning, machine learning, and deep learning with models and algorithms from the field of psychometrics. The results are interesting new approaches that can be used in psychometric practice, and that can -in turn- aid in increasing the interpretability of the black-box models used in the field of artificial intelligence.

The outline of this symposium is as follows: In the first two talks the interest is in modeling individual differences in cross-sectional data: First, **Molenaar** will study the usefulness of autoencoders as a non-parametric estimator of individual differences in multidimensional item response theory (MIRT) models. Next, **Veldkamp** will focus on the issue of missing data in these, and variational, MIRT autoencoders. In the next two talks, the focus is on longitudinal modeling: First, **Chow** will propose methods based on wavelets, like wavelet scattering, to model the change of different mood measures over time. Next, **Fokkema** will talk about the use of generalized additive model trees to detect subgroups with different non-parametric trajectories of change over time.

Autoencoders for amortized joint maximum likelihood estimation of item response theory models

Tuesday, 16th July - 14:30: Symposium: Statistical learning approaches to psychometric challenges (NB D) - Symposia

Dr. Dylan Molenaar (University of Amsterdam), Dr. Raoul Grasman (University of Amsterdam), Dr. Mariana Cúri (University of Sao Paulo)

Neural networks like variational autoencoders have been proposed as a statistical tool to fit item response theory models to data. Advantages are that high dimensional models can be estimated more efficiently as compared to conventional approaches. In this study, we demonstrate advantages of a specific autoencoder as a tool for amortized joint maximum likelihood estimation of item response theory models. Contrary to contemporary joint maximum likelihood estimation and marginal maximum likelihood estimation, no additional parameter constraints are necessary to ensure standard asymptotic theory to apply. In a simulation study, the performance of the autoencoder is compared to constrained joint maximum likelihood and various forms of marginal maximum likelihood under different distributions for the factor scores. Results show that the amortized joint maximum likelihood estimates of the factors scores are overall less biased as compared to the other approaches. We illustrate the use of the autoencoder in a real data example.

One shape may not fit all: Detecting differences in trajectories using GAM trees

Tuesday, 16th July - 14:30: Symposium: Statistical learning approaches to psychometric challenges (NB D) - Symposia

Dr. Marjolein Fokkema (Leiden University)

Modeling trajectories of change over time is of key interest in many fields of behavioral research: Clinical psychologists may want to model symptom trajectories, educational psychologists may want to model student's academic skills over time, cognitive psychologists may want to model event-related potentials.

Parametric methods such as SEMs or GLMMs are often used to model trajectories over time, but require the researcher to specify the trajectory shape a priori. Generalized additive models (GAMs) can flexibly approximate trajectory shapes through the estimation of smoothing splines. Yet, smoothing spline models still assume that the same shape fits all observations in a dataset equally well. To allow for different shapes, subgroups presenting with different shapes need to be specified by the researcher in advance.

Often, these subgroups are not known in advance and researchers may want to detect them. GAM trees allow researchers to do that, if time-constant covariates that are potentially predictive of trajectory shapes are available. This set of covariates may possibly be (very) large. GAM trees use REML for estimating the smoothing splines (Wood, 2004), and model-based recursive partitioning (Zeileis, Hothorn & Hornik, 2008) and mixed-effects model derivatives (Wang, Graves, Rosseel & Merkle, 2022) for identifying subgroups.

This presentation will explain the underlying principles of GAM trees, present empirical results on their performance and showcase a real-world application on academic trajectories.

Affect Modeling with Wavelet-Based Time Series and Statistical Learning Methods

Tuesday, 16th July - 14:30: Symposium: Statistical learning approaches to psychometric challenges (NB D) - Symposia

Dr. Sy-Miin Chow (The Pennsylvania State University), Jyotirmoy Das (The Pennsylvania State University), Yanling Li (The Pennsylvania State University), Soundar Kumara (The Pennsylvania State University)

Affective processes have been reported to show distinct changes across multiple temporal scales. The phrase "*moods nag at us, emotions scream at us*" (Larsen, 2000; p. 130) was used to clarify the distinctions between emotions, which are short–lived, relatively intense, and are triggered by specific events or targets; and moods, which reflect longer–term feelings that may not have a specific cause. The issue of multiple temporal scales is especially pertinent to areas such as affect forecasting, the prediction of how individuals feel in the future; and affective classification, the process of distinguishing or contrasting one discrete emotional state from another. We present results from using wavelets in time series analysis and statistical learning to understand the across-time and across-frequency strengths of the dynamics and lead-lag associations (contagions) across affective measures (e.g., self-reports, and physiological measures). Among these methods is wavelet scattering, a class of machine learning methods that combines, in a single step, the strengths of wavelet transform and deep neural networks to enable use of deep learning for forecasting and classification purposes with features from wavelet analysis. Because coefficients in the deep neural network are fixed to known coefficients in the wavelet analysis, computational burden and expenses are greatly reduced, with useful results found even with sample sizes that are comparably small for standard machine learning applications. The importance of hyperparameter tuning, model interpretation tools, and ways to triangulate wavelet-based methods to advance knowledge on dynamic networks and future designs of intensive longitudinal studies involving multi-scale processes are discussed.

Handling Missing Data In Variational Autoencoder based Item Response Theory

Tuesday, 16th July - 14:30: Symposium: Statistical learning approaches to psychometric challenges (NB D) - Symposia

Mr. Karel Veldkamp (University of Amsterdam)

Recently Variational Autoencoders (VAEs) have been proposed as a method to estimate high dimensional Item Response Theory (IRT) models on large datasets. Although these improve the efficiency of estimation drastically compared to traditional methods, they have no natural way to deal with missing values. In this paper, we adapt three existing methods from the VAE literature to the IRT setting and propose one new method. We compare the performance of the different VAE-based methods to each other and to marginal maximum likelihood estimation for increasing levels of missing data in a simulation study for both three- and ten-dimensional IRT models. Additionally, we demonstrate the use of the VAE-based models on an existing algebra test dataset. Results confirm that VAE-based methods are a time-efficient alternative to marginal maximum likelihood, but that a larger number of importance-weighted samples are needed when the proportion of missing values is large.

Group-specific detection rates when using tree-based differential item functioning methods

Tuesday, 16th July - 14:30: Differential Item Functioning (RB 209) - Oral

Prof. Ronna Turner (University of Arkansas), Nana Amma Asamoah (University of Arkansas)

Tree-based approaches have advantages for differential item functioning (DIF) assessment as recursive partitioning methods do not require *a priori* cutoffs on continuous variables and multiple covariates can be combined in single data analyses. In this study, we compare two tree-based DIF approaches: Rasch trees (Strobl et al., 2015) and item-focused trees (Tutz & Berger, 2018). These methods use different approaches for item parameter estimation and DIF evaluation and have different true and false positive rates (TPR, FPR) under varying conditions. For example, item-focused trees have higher DIF detection when DIF occurs in a single item, whereas Rasch trees may be more effective at identifying small DIF favoring the same group in multiple items (Bollman et al, 2018). These procedures have been evaluated for categorical and continuous covariates, with a newer study investigating performance when the proportion of items favoring each group are equal (balanced) versus unequal (unbalanced) (Asamoah et al, 2022). However, prior studies are limited to evaluating TPR and FPR for samples as a whole. Research with other DIF methods has indicated that TPR and FPR differ for subgroups when the proportion of DIF items favoring each group is unequal (Turner & Keiffer, 2019). Similarly, disproportionate DIF items are likely to impact parameter and DIF estimation when using tree-based methods. This study will investigate how balanced and unbalanced DIF designs impact group-specific DIF detection under varying data conditions and compare these impacts across the two tree-based methods. Results will inform researchers on when purification may be recommended.

Using criterion validity to solve the scale indeterminacy in the context of DIF: An external variable approach

Tuesday, 16th July - 14:45: Differential Item Functioning (RB 209) - Oral

Dr. Jesper Tijmstra (Tilburg University), Dr. Maria Bolsinova (Tilburg University)

An important fundamental issue when evaluating measurement invariance in IRT is the establishment of a common scale for the considered groups. This is commonly solved by assuming that the scales for the different groups can be linked through a DIF-free anchor set of items, and subsequently the rest of the items can be tested for DIF. However, with a priori knowledge about items being DIF-free being questionable, obtaining a trustworthy anchor set is often problematic (or outright impossible, if all items have DIF). A related solution which assumes that the majority of items are DIF-free is also not always justifiable.

We propose an alternative approach to both solving the scale indeterminacy in the presence of DIF, and to DIF-testing itself, which does not rely on assumptions about DIF-free items being present. Instead we propose an external variable approach, where a relevant criterion variable is used to solve the indeterminacy of the scale linking. By selecting an external variable that is known to have a high correlation with the latent variable to be measured, and by optimizing the scale linking to maximize the correlation between the latent variable and the criterion variable, a scale linkage can be found that maximizes predictive validity. With this link established, item-level DIF tests can subsequently also be performed, allowing for anchor-free DIF testing.

In this presentation we will discuss the proposed method, and illustrate its feasibility and relevance through simulation studies. Both theoretical and practical consequences of the approach will be discussed.

Calculating bias in equating and handling DIF anchor items

Tuesday, 16th July - 15:00: Differential Item Functioning (RB 209) - Oral

Prof. Marie Wiberg (Umeå university), Inga Laukaityte (Umeå university)

Test score equating is used to make scores from different test forms comparable, and the nonequivalent group with anchor test design is commonly used. In standardized tests, we do not want items with differential item functioning (DIF), i.e. when groups with the same latent ability but from different groups have an unequal probability of answering a given item. DIF anchor items may impact the equating transformation. The overall aim was to compare chained equating and frequency estimation when we have DIF in the anchor test using the two evaluation measures: standard error of equating and bias. We used simulated and real test data from the Swedish Scholastic Aptitude Test (SweSAT), which is a multiple-choice binary scored test used for college admissions. In the simulations, we examined DIF in the anchor test, difference in ability groups, and difference in item difficulty. Maentel-Hanszel was used to detect DIF items. Bias can be calculated either by equating the test forms to itself or use a criterion function. In this study we examined bias both when equating the test forms to themselves and four different criterion functions: the identity function, equipercentile, chained equating and frequency estimation. The results indicate that the method of bias calculation significantly influences the determination of appropriate equating methods for various scenarios. Practical implications for how to handle if we have DIF items in the anchor test form in standardized tests are given together with recommendations on how to calculate bias when evaluating equating transformations.

Addressing heterogeneity in multisubject multivariate repeated measures data

Tuesday, 16th July - 14:30: Symposium: Heterogeneity in Repeated Measures (RB 210) - Symposium Overview

Prof. Timothy Brick (The Pennsylvania State University)

In the evolving landscape of scientific research, the recognition of heterogeneity across people represents a pivotal shift in how we understand and analyze the complexity of real-world phenomena. Here we present four novel analytical strategies to address heterogeneity in complex multisubject multivariate repeated measures data. The first is the introduction of SEM-Boruta, an adaptation of the Boruta feature-selection algorithm that integrates Structural Equation Model (SEM) Forests (Brandmaier et al., 2016). Before testing models of process, this algorithm proves beneficial to identify relevant moderators from complex datasets. Our second proposal is an “idiothetic” method built from the idiographic framework’s limitations in generalizability of person-specific results to broad classes of individuals. This approach is a continuous-time extension of the group iterative multiple model estimation (GIMME; Gates & Molenaar, 2012) procedure, offering a refined modeling of individual processes and group-level structures. Third, we present the multi-VAR framework (Fisher et al., 2022, 2024) for simultaneously estimating common, individual, and partially shared dynamic features from multiple-subject time series data. This approach uses structured penalties to share information across individual-level models and improves upon existing methods for accommodating both quantitative and qualitative differences in model dynamics. Lastly, the fusion of state space models with mixture models, as explored by Hunter (2014, 2024), advances the understanding of individual change over time by accounting for qualitatively distinct patterns in both measurement and dynamics. These innovations represent significant strides in statistical methodologies, enabling an enhanced understanding of individual and group-level heterogeneity in biobehavioral research.

Improving Boruta feature selection in multivariate framework

Tuesday, 16th July - 14:30: Symposium: Heterogeneity in Repeated Measures (RB 210) - Symposia

Ms. Priyanka Paul (The Pennsylvania State University), Prof. Timothy Brick (The Pennsylvania State University), Prof. Andreas Brandmaier (Max Planck Institute for Human Development Berlin)

With increasing access to large-scale datasets, it is easier to create and test models of process, but identifying important moderators in a sea of possible influential variables can be quite challenging. Furthermore, the presence of heterogeneity in data necessitates advances in analytical strategies. We introduce SEM-Boruta, a novel adaptation of the Boruta feature selection algorithm, and explore its robustness through application on simulated datasets within a multilevel multigroup framework. Boruta utilizes Random Forest Classifiers, a machine learning procedure based on recursive partitioning that is often used to identify potentially important predictors. By adapting Boruta to use Structural Equation Model (SEM) Trees (Brandmaier et al., 2013), we convert it from a *predictor* selection algorithm into a *moderator* selection algorithm: a guided form of heterogeneity search. We present initial validation of SEM-Boruta's robustness and provide insights into its ability to discern relevant features in complex data structures. This investigation also underscores the importance of tools like SEM-Boruta in identifying heterogeneous subsets in a population and provides a nuanced understanding of heterogeneity in research data.

Are General Truths Generally True? Accommodating Meaningful Heterogeneity in Dynamic Processes

Tuesday, 16th July - 14:30: Symposium: Heterogeneity in Repeated Measures (RB 210) - Symposia

Dr. Zachary Fisher (The Pennsylvania State University), Mr. Christopher Crawford (The Pennsylvania State University), Dr. Vladas Pipiras (University of North Carolina - Chapel Hill)

In this talk we present recent developments in the use of structured penalization and estimation for accommodating heterogeneous dynamics in multiple-subject multivariate time series. Specifically, we discuss the multi-VAR framework as described by Fisher et al. (2022,2024) for modeling subject-specific time series models with common and individualizing features. This approach differs from many popular methodologies for multiple-subject time series in that both qualitative and quantitative differences in a large number of individual-level dynamics are well-accommodated. In this talk novel extensions to the multi-VAR framework will be presented. These extensions include the identification of subgroup-level effects, fusion penalties designed to capture partially shared dynamic structures, and the integration of discrete-valued marginal distributions for the component series.

Making extremely heterogeneous person-specific dynamics so happy together: State space mixtures all the way down

Tuesday, 16th July - 14:30: Symposium: Heterogeneity in Repeated Measures (RB 210) - Symposia

Prof. Michael Hunter (The Pennsylvania State University)

Repeated measures data are not new, yet statistical methods are only recently being developed to fully capitalize on these rich sources of information to better understand how people change over time. A true science of the individual must confront the fundamental challenge that people are often far more heterogeneous than is assumed by conventional models, even many modeling approaches that claim to account for heterogeneity. More fully utilizing multisubject time series data to understand foundational questions of how people change over time is the purpose of the present work. Hunter (2014, 2024) showed previously that a kind of latent time series model called a state space model can be combined with mixture models to account for qualitatively distinct patterns of dynamics under the assumption of equivalent measurement structures across people. This combination is the state space mixture model. The present talk evaluates the expansion of these capabilities to two novel cases. First, we investigate the converse problem: can we allow for qualitatively distinct patterns of measurement under the assumption of equivalent dynamics? Second, we allow for qualitative differences in both measurement and dynamics. People may have different factor structures or even dimensions along with distinct dynamics. Simulated data are used to graphically and numerically demonstrate both of these cases, along with a small comparison to some multilevel modeling analogs. If true differences between people are as severe as the simulated situations, multilevel modeling is insufficient to account for them.

Unsupervised model construction in continuous-time

Tuesday, 16th July - 14:30: Symposium: Heterogeneity in Repeated Measures (RB 210) - Symposia

Jonathan Park (University of California, Davis), Dr. Sy-Miin Chow (The Pennsylvania State University), Dr. Zachary Fisher (The Pennsylvania State University), Dr. Michael Hunter (The Pennsylvania State University), Dr. Peter Molenaar (The Pennsylvania State University)

The idiographic framework has been invaluable for understanding the dynamics of individuals as they unfold through time. However, some concerns have been levied regarding the generalizability of person-specific results to broad classes of individuals. Addressing the challenge of bridging person- and group-level inference remains a pivotal issue in quantitative methods. A broad class of models—referred to as “idio-thetic” methods—describe the spectrum between purely idiographic models to fully constrained group-level or “chained” models. Much of the advancement in idio-thetic methods has been within the realm of discrete-time literature. Discrete-time models offer intuitive interpretations and ease of implementation; however, they are limited when compared to continuous-time models. The latter are adept at handling unequally spaced data—which frequently occur in real-world empirical applications—and possess other unique advantages. Despite this, development of novel methods in continuous-time still lags behind those in the discrete-time literature. I introduce a continuous-time extension of the group iterative multiple model estimation (GIMME; Gates & Molenaar, 2012) procedure. This work streamlines the fitting of continuous-time models to individual processes and leverages person-specific information to identify common, group-level structures. The discussion will cover the formal algorithm and simulation results which highlight the strengths of moving into a continuous-time framework.

Developing cross-classified growth mixture models to account for student mobility

Tuesday, 16th July - 14:30: Topics in Longitudinal Data Analysis (RB 211) - Oral

Dr. Yan Wang (University of Massachusetts Lowell), Tonghui Xu (University of Massachusetts Lowell), Dr. Nilabja Guha (University of Massachusetts Lowell), Dr. Audrey Leroux (Georgia State University), Alexandria Winstead (University of Massachusetts Lowell), Christopher Simmons (University of Massachusetts Lowell), Dr. Chunhua Cao (University of Alabama)

Growth mixture modeling (GMM) is an increasingly popular longitudinal data analytic approach to identifying subpopulations of students that differ in their growth trajectories of academic outcomes. Given that educational data often have a hierarchical structure in which, for example, students are nested within schools, GMM has been extended to multilevel GMM that takes such nested data structure into account (Muthén & Asparouhov, 2009; Palardy & Vermunt, 2010). However, current methodological techniques of multilevel GMM do not allow researchers to account for student mobility which is a prevalent and critical social justice issue given its predominant impact on minority student populations and their academic achievements. Student mobility has been ignored by not considering school contexts at all, deleting mobile students, or choosing one school membership for mobile students. To address this methodological limitation, this study proposes a cross-classified GMM (CC-GMM), which accounts for student mobility with repeated measures cross-classified by students and schools. Because students and schools are viewed as separate higher levels of analysis (rather than students nested within schools), the proposed model allows for multiple school memberships over time (e.g., Chen & Leroux, 2018; Grady & Beretvas, 2010; Luo & Kwok, 2012). Random effects due to students and schools can be decomposed and estimated. The broad utility of CC-GMM will be illustrated with the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 data. This study contributes to the methodological advancement of GMM and finite mixture modeling in general. Implications on educational, social, and behavior sciences will be discussed.

A New Approach of Integrating Nonlinear Growth Modeling on Accelerated Longitudinal Design

Tuesday, 16th July - 14:45: Topics in Longitudinal Data Analysis (RB 211) - Oral

Ms. HeeJoo Park (Yonsei University), Prof. Ji Hoon Ryoo (Yonsei University)

Longitudinal studies aim to identify how characteristics or trajectories change across time. In particular, the purpose of age-related longitudinal research is to estimate changes in growth over age span. However, such a pattern of change is often represented as nonlinear growth rather than a linear growth, especially for the longer age span. Nonlinear growth models can estimate slopes, inflection points, rates of change, initial levels, and change points, which enables to understand various aspects of development in longitudinal studies. Those characteristics of nonlinear growth are often represented by global functional forms such as conventional polynomial functions, exponential and logarithm functions, Gompertz, etc as well as piecewise growth models such as splines.

Accelerated longitudinal design (ALD) is a methodological tool to extend an extrapolative range of a cohort by utilizing data of multiple cohorts collected within a same study period in a longitudinal study. In this study, we propose an additional nonlinear method by applying nonlinear growth modeling within the ALD framework. Also, we introduce nonlinear growth modeling within ALD framework focusing on fractional polynomials for the ALD to remedy some issues in longitudinal study, which also allows researchers to expand flexibility of growth models. We demonstrate its efficiency of utilizing parameters estimated in practice via empirical data analysis, student's performance data in mathematics education, and we also conducted a simulation study to confirm the results. Results show flexibility of nonlinear modeling by using fractional polynomials, efficiency compared with linear or conventional polynomial models, and predictability of ALD models obtained.

Validation of the intra-individual speed-ability relationship (ISAR) model

Tuesday, 16th July - 15:00: Topics in Longitudinal Data Analysis (RB 211) - Oral

Augustin Mutak (Freie Universität Berlin), Sören Much (Martin-Luther-Universität Halle-Wittenberg), Dr. Jochen Ranger (Martin-Luther-Universität Halle-Wittenberg), Prof. Steffi Pohl (Freie Universität Berlin)

The speed-ability trade-off (SAT) is extensively documented in the psychological literature, illustrating a decline in task performance due to an increase in speed. Recently, a new psychometric model, the ISAR model (Mutak et al., in press), has been developed to capture the SAT. This model extends van der Linden's (2007) hierarchical speed-accuracy model by incorporating growth terms for speed and ability. Consequently, it enables the exploration of the correlation between intra-individual changes in speed and ability, which is thought to reflect the speed-ability trade-off. However, attributing this correlation solely to the SAT can be problematic, as other factors such as concentration and motivation may also change during the test, and confound the intra-individual relationship ability and speed. To disentangle the different effects, we conducted an empirical study comprising a matrix reasoning test as the primary measure and concentration and motivation as confounding measures assessed at various time points. In analyses we control for changes in concentration and motivation. We evaluate not only their impact on the intra-individual speed-ability relation but also aim at getting a more informative measure of the SAT after controlling for these variables. The results aim at validating the interpretation of the model parameters and may also guide future work on other approaches that rely on non-stationarity of speed and ability when investigating the SAT.

Joint Modeling: Dynamic Latent Variables, Survival Outcomes, and MNAR Data

Tuesday, 16th July - 15:15: Topics in Longitudinal Data Analysis (RB 211) - Oral

Ms. Yvette Baurne (Department of Statistics, Lund University)

Within the social and behavioral sciences, a substantive research focus is the interplay between latent dynamic variables, significant events and their characteristics, and decision making. As an example, one may be interested in the relations between mental well-being, an event and how traumatizing it is perceived as, and the decision to seek professional help. For this type of research questions we use longitudinal data, often with ordinal responses. Longitudinal data frequently comes with the challenge of missing data, both intermittent and drop-out. Separately these questions and challenges are well studied and several methods exists to address them, but when interest lies in dealing with them all at once the task is more challenging. Within a Bayesian framework we introduce a model that jointly models dynamic latent variables with ordinal outcomes using a latent Gaussian model approach, decision making in terms of a survival outcome using a Weibull hazard function, and a selection model to account for data missing not at random (MNAR). We apply the method to a research question of individual team members' trust in their team and how it develops over time, how it is related to the perceived novelty and disruptiveness of events, and the consequences for team member exit from the team.

Theory Construction Methods

Tuesday, 16th July - 14:30: Symposium: Theory Construction Methods (RB 212) - Symposium Overview

Mr. Jason Nak (University of Amsterdam)

In this symposium we will present a collection of current developments in methodologies for theory construction. Although methodological development in psychological science has been mostly focussed on the empirical, there has been an increasingly stronger call for theory construction and several lines of methodological development have taken off. We would like to present several avenues of theory construction that are currently being pursued. Rasoul Norouzi Nikjeh will present on the use of AI models in navigating the ever-expanding body of knowledge to extract causal claims, discussing development on the “Text Mining Systematic Reviews” framework. Meike Waaijers will present her new R-package which helps researchers create causal loop diagrams containing multiple nodes (e.g. symptoms, constructs, variables) within a topic of interest by querying large language models. Adam Finneman will give a talk about using statistical mechanics to construct theories about networks of interacting elements. As many psychological constructs are believed to operate as such networks, we may be able to use equivalent models from other fields to explain them. Han van der Maas will give a presentation on the Blume-Capel spin model. This model extends the commonly used Ising model by adding neutral states, enabling the modeling of more complex dynamical systems. Finally, Noah van Dongen and Jason Nak will present the development of parallel databases for the collection of phenomena, stable features which are suitable for use as building blocks in theories, and formal models, computational or mathematical representations which can be used to formalize psychological theories.

Text mining systematic reviews: New directions for qualitative research synthesis

Tuesday, 16th July - 14:30: Symposium: Theory Construction Methods (RB 212) - Symposia

Mr. Rasoul Norouzi (PhD Student of Text Mining Research Synthesis Methods, dept. Methodology & Statistics, Tilburg University, Netherlands), Dr. Caspar van Lissa (Associate Professor of Social Data Science, Dept. Methodology & Statistics, Tilburg University.), Dr. Bennett Kleinberg (Associate Professor of Behavioural Data Science, Dept. Methodology & Statistics, Tilburg, Netherlands.)

As the number of publications in most fields continues to grow exponentially, it becomes increasingly unfeasible for scholars to remain informed about the entire literature. Moreover, narrative reviews are subjective and susceptible to a variety of biases, including confirmation bias. Innovations in the machine learning domain of text mining can be used to synthesize the burgeoning literature in a relatively objective and scalable manner. This presentation discusses pioneering work on Text Mining Systematic Reviews (TMSR, Van Lissa 2022). TMSR is an umbrella term for quantitatively aided qualitative research synthesis methods that use machine learning to extract knowledge from published scientific literature. I present two studies that used different TMSR pipelines to extract a latent nomological network from the literature in a particular subfield. These networks can serve as a useful starting point for theory development, help researchers find their bearings in the literature, and identify knowledge gaps. Furthermore, I present one application of TMSR that sets out to identify causal claims in the literature. In fields where causal assumptions are rarely made explicit (especially social science), extracting the latent causal network from the published literature may advance a more explicit discussion about causality and formal theory.

AI-assisted theory construction

Tuesday, 16th July - 14:30: Symposium: Theory Construction Methods (RB 212) - Symposia

Meike Waaijers (University of Amsterdam), Hannes Rosenbusch (University of Amsterdam), Prof. Denny Borsboom (University of Amsterdam)

Large Language Models (LLMs) excel at processing and constructing sentences in natural language and can analyse vast amounts of information, enabling them to recognise patterns and draw conclusions that may be elusive to human researchers. Their ability to interpret and follow detailed textual instructions enables LLMs to follow complex theoretical modelling tasks, such as the causal loop diagram (CLD) method. The CLD method involves a collaborative process in which a focus group of experts use their collective knowledge to identify and map out relationships between variables in a system, highlighting feedback loops and causal connections. We argue that LLMs can accelerate and enrich this theory construction process by identifying connections and insights that may remain undetectable to researchers. To address difficulties in theory construction in psychological research, we have developed an R package that uses LLMs to generate candidate CLDs. The package includes seven functions that allow users to visually represent key variables and their causal relationships in any research domain. This R package demonstrates the potential of LLMs to aid theory construction in psychology, providing a quick and efficient way to develop robust candidate theories.

A theory construction framework for networks

Tuesday, 16th July - 14:30: Symposium: Theory Construction Methods (RB 212) - Symposia

Mx. Adam Finnemann (University of Amsterdam)

In this talk, I will present a framework for constructing theories about networks of interacting elements - a common structure believed to underlie psychological constructs such as personalities, emotions, decisions, and psychopathologies. We can study such systems through formal models in order to derive novel integrative explanations of non-network phenomena. The Ising model is the most commonly used model for studying such systems due to its simple structure yet rich dynamics. However, we demonstrate that it is just one of six foundational models that together form a unified framework for theory construction. Additionally, we show how these foundational models can be expanded to create a versatile framework. To analyze and validate the constructed models, we present three methods. Finally, the limitations of this framework are discussed, highlighting areas for future research and improvement.

PsychoModels and PsychoFacts: A Platform for Formal Theory Development in Psychological Science

Tuesday, 16th July - 14:30: Symposium: Theory Construction Methods (RB 212) - Symposia

Dr. Noah van Dongen (University of Amsterdam), Mr. Jason Nak (University of Amsterdam)

Beyond improving empirical methods, there have been many calls for developing stronger theories in psychological science. If psychology is to advance, both a unified knowledge base and more concrete theoretical underpinnings are required. To assist in this endeavor, we are developing two databases: one storing established psychological phenomena, the other storing computational and mathematical models created for psychological science.

Through the development of these databases, we aim to create a platform where researchers can create formal theories through selecting, matching, combining, and adapting the models and phenomena collected in the databases. In the *PsychoFacts* database, researchers will find phenomena and the statistical patterns that they imply (e.g., a response-time delay in the Stroop-effect), which require explanation by psychological theory. In the *PsychoModels* database, researchers can store and share computational and mathematical models, which can represent relevant mechanisms in psychological theory. Ultimately, researchers should be able to evaluate theories by matching simulated data from (adjusted) models selected from the *PsychoModels* database to the statistical patterns of the relevant phenomena selected from the *PsychoFacts* database.

Our talk outlines the architecture of this platform and the ontologies of the two databases. We will present work-in-progress versions of *PsychoFacts* and *PsychoModels* and demonstrate how the platform could be used by potential theorists. We hope that our efforts can benefit theoretical and empirical work, aid didactic efforts around modeling, and expand our perspective on psychological research.

Constructs May or May Not Be Latent: Studies on Two Domains of Structural Equation Modeling

Tuesday, 16th July - 16:15: Dissertation Award Presentation (Vencovského aula) - Dissertation Award Talk

Dr. Gyeongcheol Cho (Department of Psychology, The Ohio State University)

Structural equation modeling (SEM) enables empirical testing of hypothetical relationships between variables, including constructs. Traditionally, SEM has assumed all constructs to be latent, existing independently of their indicators, and represents them by a (common) factor. This domain is known as factor-based SEM. However, some constructs, such as socioeconomic status and genes, are not inherently latent but rather correspond to a summary or cluster of their indicators. To deal with such constructs, component-based SEM has emerged, wherein constructs are represented as composite indexes of indicators, termed components.

In this talk, I begin with a systematic comparison of the two SEM domains and briefly introduce three novel SEM methods—structured factor analysis (SFA), convex generalized structured component analysis (convex GSCA), and deep learning generalized structured component analysis (DL-GSCA). SFA tackles two long-standing issues in factor-based SEM—improper solutions and factor score indeterminacy—by simultaneously estimating model parameters and the probability distribution of (candidate) factor scores using a single cost function. Conversely, convex GSCA and DL-GSCA extend the capabilities of component-based SEM; the former generates interpretable components that preserve the original scales of indicators, while the latter employs deep learning to identify non-linear components that maximize predictive power for target outcome variables. I will demonstrate the technical foundations of these methods and their practically utility through empirical data analyses.

Foundational Competencies in Educational Measurement: An NCME Task Force Consensus

Wednesday, 17th July - 09:00: Invited Panel: Foundational Competencies in Educational Measurement: An NCME Task Force Consensus (Vencovského aula) - Invited Panel

Prof. Andrew Ho (Harvard University)

I present the published consensus of a National Council on Measurement in Education (NCME) Presidential Task Force on Foundational Competencies in Educational Measurement and provide additional unpublished commentary on its intersections with quantitative psychology. Foundational competencies are those that support future development of additional professional and disciplinary competencies. The authors reviewed job postings, course syllabi, and classroom activities to develop a framework for foundational competencies in educational measurement.

The framework introduces three foundational competency domains: 1) Communication and Collaboration Competencies; 2) Technical, Statistical, and Computational Competencies; and 3) Educational Measurement Competencies. Within the Educational Measurement Competency domain, the authors identify five subdomains: 3A) Social, Cultural, Historical, and Political Context; 3B) Validity, Validation, and Fairness; 3C) Theory and Instrumentation; 3D) Precision and Generalization; and 3E) Psychometric Modeling. The full article is available here and includes a glossary and illustrated framework in Figures 1 and 2: https://onlinelibrary.wiley.com/doi/abs/10.1111/emip.12581

The remainder of my presentation compares and contrasts this framework with a counterfactual question: What might the framework have been, had there been a Psychometric Society Task Force for Foundational Competencies in Quantitative Psychology? I suggest that there should be many contrasts, and I list three here: 1) Psychological subdomains would be clearer in the contrasting framework. 2) The role of context would remain important, including historical development and international contexts of applications of quantitative psychology. 3) Causal elements underlying quantification and measurement would be more salient, including the role of constructs in causing item responses and the role of context in moderating item responses.

Using process data to understand, assist, and enhance assessment

Wednesday, 17th July - 09:00: Symposium: Using process data to understand, assist, and enhance assessment (RB 101) - Symposium Overview

Dr. Mengxiao Zhu (University of Science and Technology of China)

Technology enables collection of process data from computer-based assessments conveniently. Through system logs or external devices, such as eye-trackers, process data record the time stamps and related actions of the learners during assessment, which can be used to recover learners' response processes during the assessment that go beyond the final responses.

This symposium includes five studies focusing on different ways of leveraging the power of process data to understand, assist and enhance assessment. The first study explores the linkage of writing process data to the quality of argumentative writing tasks through the analysis of keystroke logs. The second study identifies a set of sequence features from process data as latent class mediators to account for the observed performance gap between learners with and without learning disabilities. The third study builds deep learning models to predict the scores of argumentative writing essays using both keystroke process data and writing content information. The fourth study collects eye-tracking data during reading processes and adopts ensemble learning methods to build classifiers for identification and early screening of reading disorders in children. The last study analyzes log files from a scenario-based scientific inquiry task for a fine-grained diagnosis of students' scientific inquiry learning using a cognitive diagnostic model.

Altogether, this symposium showcases innovative ways of using the rich information contained in the process data to deepen our understanding of the assessment responses and scores, to explain the differences in behavior patterns among test takers, and to provide evidence for assessment and diagnostics.

Linking Writing Process Data to Writing Quality

Wednesday, 17th July - 09:00: Symposium: Using process data to understand, assist, and enhance assessment (RB 101) - Symposia

Hong Jiao (University of Maryland), Chandramani Lnu (University of Maryland), Nan Zhang (University of Maryland), Xiaoming Zhai (University of Georgia)

Process data may add value to assessment. This study explores the linkage of writing process data to the writing quality of argumentative writing tasks. Four SAT writing prompts were used to collect the keystroke data and writing scores for 5000 participants randomly assigned to each prompt, scored on a scale of 0 to 6. The instructions required the minimum essay length of 200 words in 3 paragraphs. A keystroke logging program written in vanilla Javascript embedded in the script of the website. The program listened to the keystroke and mouse events. The time stamp and cursor position for each keystroke or mouse operation was logged. The identified operations include input, delete, paste, and replace and text changes. The dataset only include the information about the text change, but no information was available about the exact text input or text changes. Keystroke measures extracted from the log data included production rate, the essay length, average word and sentence length, total response time, pause, warm-up time, revision types and frequencies, and burst. Different base models were compared including LightGBM-Regressor, CatBoostRegressor, SVR, and XGBRegressor. VotingRegressor was used to ensemble the model outputs. Both root mean squared error of the predicted scores and QWK were computed to quantify the prediction accuracy. The ensemble model yielded the smallest RMSE of 0.5861 while LightGBM-Regressor yielded the highest QWK of 0.76. Feature importance was examined and the following features turned out to be the top important features: sentence length, revision counts, word length, pause time, paragraph length.

Explaining performance gaps with problem-solving process data via LCMA

Wednesday, 17th July - 09:00: Symposium: Using process data to understand, assist, and enhance assessment (RB 101) - Symposia

Sunbeom Kwon (University of Illinois, Urbana-Champaign), Prof. Susu Zhang (University of Illinois, Urbana-Champaign)

The use of computers as assessment delivery platforms allowed the collection of process data documenting an examinees' sequence of actions in pursuit of solving a task (e.g., clicks, keystrokes, and revisits). Such data can provide valuable information on how examinees arrived at their final outcome, and thus, can provide information beyond response data. Analyzing process data can aid in understanding subgroup differences (e.g., He & von Davier, 2016; Liao, He, & Jiao, 2019) and explaining differences in sequential patterns in correct/incorrect problem-solving (e.g., Greiff, Wüstenberg, & Avvisati, 2015; Ulitzsch, He, & Pohl, 2022).

In this study, we propose a latent class mediation analysis procedure to reveal the latent classes underlying the distribution of sequence features extracted from the process data. This approach aims to explain the performance gap between groups in the outcome. Employing a variable selection algorithm, our approach identifies the optimal set of sequence features, resulting in latent class mediators that effectively account for the observed performance gap.

We analyze process data collected from the National Assessment of Educational Progress (NAEP) math assessment. The NAEP math assessment results showed that learners with learning disabilities (LD) continuously exhibit poorer math performance than their typically developing (TD) peers without disabilities. The proposed analysis procedure allows explaining the achievement gap by uncovering latent classes related to the test-taking process leading to the difference in performance between the two groups. Our research contributes to the broader discussion within psychometrics on using process data to enhance the fairness and equity of educational assessment.

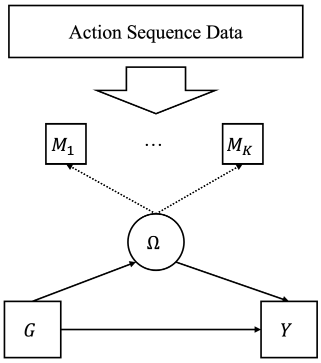


Figure 1. Latent Class Mediation Analysis. M_k is a process feature, Ω is a latent class variable, G is a binary group membership variable, and Y is a binary outcome variable.

Figure1.png

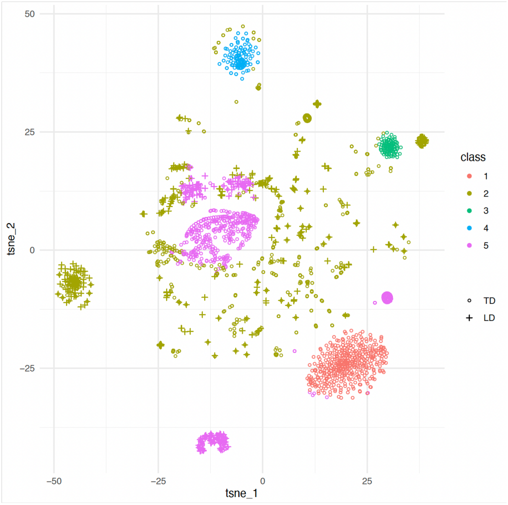


Figure 2. t-SNE plot for NAEP math assessment data. Five latent classes were identified.

Figure2.png

Scoring of argumentative writing using keystroke logs and deep learning models

Wednesday, 17th July - 09:00: Symposium: Using process data to understand, assist, and enhance assessment (RB 101) - Symposia

Dr. Mengxiao Zhu (University of Science and Technology of China), Qi Shu (University of Science and Technology of China), Mo Zhang (Educational Testing Service), Dr. Qizhi Xu (University of Science and Technology of China)

To assess writing skills, the learners are often asked to prepare essays or reports, which are often evaluated by either human raters or automated essay scoring systems. In addition to the final writing products, keystroke logging records all keyboarding events and the corresponding timestamps using computer system logs, which contains valuable information and has the potential to be used for assessment of writing and for providing feedback for improvement.

This study uses writing keystroke logs from a scenario-based assessment (SBA) of writing, collected as part of a larger data collection of middle-school students participating in the CBAL Writing Study (van Rijn et al., 2016). Using the SBA design, before assigning the writing tasks, participants were prompted with scenarios and source reading materials of a unifying theme. The keystroke logs include both the writing actions, such as insert, delete and replace, and the writing contents. A total number of six prompts were used in this study with about 700 to 1,100 participants for each prompt. All essays were graded on two rubrics each ranging from 0 to 5, with the first one on discourse features and the second one on argumentative writing features. Preliminary results revealed the effectiveness of automated scoring through deep learning models, such as BERT, Transformer, and LLM, on both rubric scores using the keystroke logs and the writing content on the argumentative writing essays. Further studies will explore the scoring using partial data and potential early interventions for learning of argumentative writing skills.

Selecting and analyzing eye movement features for the screening of reading disorders in children

Wednesday, 17th July - 09:00: Symposium: Using process data to understand, assist, and enhance assessment (RB 101) - Symposia

Ms. Mingming Hu (iflytek)

Large-scale early screening of reading disorders is necessary; however, the methods of traditional screening are expensive and time-consuming. With the continuous innovation and development of language research techniques, the eye-tracking-based method is introduced for the early screening of reading disorders in a text reading test. Additionally, artificial intelligence technologies, such as machine learning and deep learning, have also been gradually applied to the assessment of children's language and reading abilities because of their powerful data processing and mining capabilities. This study explores an eye movement-based method by using machine learning for the early screening of reading disorders to reduce cost and increase efficiency. In the experiments, we used a self-built eye movement dataset containing 35 children from grades three to five (26 children with high reading achievement and 9 children with low reading achievement). In addition to the traditional statistical features of eye movements in individual reading, the velocity and spatial features are also extracted. An early screening model based on eye tracking data is built through a feature selection method. The screening model achieves an accuracy rate of 84.21% in identifying children with risks of reading disorders using the ensemble learning methods to integrate multiple models, including Transformer, SVM and Logistic regression. Moreover, the top ten eye movement features that contribute most to classification are selected by using logistic regression method.

Latent variable models for complex data structures

Wednesday, 17th July - 09:00: Symposium: Latent variable models for complex data structures (NB A) - Symposium Overview

Mariagiulia Matteucci (University of Bologna)

Latent Variable Models (LVM) represent a cornerstone in statistical analysis, offering a powerful framework for modelling complex data structures and uncovering underlying relationships. However, the estimation of LVM poses numerous challenges, ranging from computational complexity with high-dimensional data to tackling specific tasks. This symposium seeks to address some issues in LVM estimation and data-related aspects across different domains by exploring innovative solutions. The five presentations will elaborate on how important challenges in modelling LVM with complex data structures can be addressed. The first talk focuses on jointly modelling grades, times, and outcomes of university students dealing with censored data within the Item Response Theory (IRT) approach. In the second work, a new method for the cheating detection in Computerized Adaptive Testing (CAT) based on response times is proposed. The third contribution deals with a modification of the factor mixture model to identify aberrant response behaviours. The fourth talk aims at comparing estimation methods of LVM with panel data, by focusing on recent approaches, such as composite likelihood methods and dimension-wise quadrature. The last work introduces a new M-estimator for Generalized Linear Latent Variable Models (GLLVM) computed by using a stochastic approximation algorithm that has favourable properties, compared to the classical estimation methods, with high-dimensional data. Overall, through a combination of simulation studies and empirical applications, the contributions in this symposium demonstrate how advanced LVM estimation and modelling techniques can be used to address practical issues in the field of psychological, educational, and social measurement.

Joint modelling of students' academic performance and dropout

Wednesday, 17th July - 09:00: Symposium: Latent variable models for complex data structures (NB A) - Symposia

Prof. Michela Battauz (University of Udine), Dr. Giuseppe Alfonzetti (University of Udine), Prof. Ruggero Bellio (University of Udine)

Academic careers are frequently analyzed using competing risk models, which relate the hazard of several alternative events to some covariates. In the context of students' academic careers, these events are graduation, dropout or transfer to another course. Besides some observed covariates, a determinant of academic careers is certainly the performance on the exams. In the Italian educational system, university students have great flexibility in the choice of the order and the time when to attempt the exams. Hence, we assume that two latent variables, ability and speed, underlie the grades and the time when the exams are attempted, which are modelled within the item response theory framework. Since these latent variables also affect the probabilities of experiencing an event of the academic career, we propose a joint model for the grades, the times and the outcomes of the academic career. In this model, there are various sources of censoring which are taken into account: the grades and the times are not observed if a student drops out or changes course before attempting the exam, the grades and the times are observed only if the exams are passed and some students might be still enrolled at the time of collecting the data.

Cheating detection based on response times in CAT

Wednesday, 17th July - 09:00: Symposium: Latent variable models for complex data structures (NB A) - Symposia

Luca Bungaro (University of Bologna), Mariagiulia Matteucci (University of Bologna), Prof. Stefania Mignani (University of Bologna), Prof. Bernard P. Veldkamp (University of Twente)

In the field of educational and psychological measurement, the shift from paper-based to computerized tests has become a prominent trend in recent years. Computerized tests allow for more complex and personalized test administration procedures, like Computerized Adaptive Testing (CAT). CAT, following the Item Response Theory (IRT) models, dynamically generates tests based on test-taker responses, driven by complex statistical algorithms. Even if CAT structures are complex, they are flexible and convenient, but concerns about test security should be addressed. Frequent item administration can lead to item exposure and cheating, necessitating preventive and diagnostic measures. We propose a method called "CHeater identification using Interim Person fit Statistic" (CHIPS), designed to identify and limit cheaters in real-time during test administration. CHIPS utilizes response times (RTs) to calculate an Interim Person fit Statistic (IPS), allowing for on-the-fly intervention using a more secret item bank. Also, a slight modification is proposed to overcome situations with constant speed, called Modified-CHIPS (M-CHIPS). The results show that CHIPS is able to correctly identify cheaters and limit the impact of cheating. In fact, the method shows good results in terms of BIAS and RMSE of the ability for cheaters. However, it struggles when cheaters answer all items correctly, a challenge addressed by M-CHIPS. The method remains robust against variations in cheaters' abilities and test-taking speeds, and it offers flexibility in test design. Limits and future developments are finally discussed.

A new factor mixture model to detect aberrant respondents

Wednesday, 17th July - 09:00: Symposium: Latent variable models for complex data structures (NB A) - Symposia

Niccolò Cao (University of Bologna), Prof. Livio Finos (University of Padua), Prof. Luigi Lombardi (University of Trento), Prof. Antonio Calcagnì (University of Padua)

In applied contexts, surveys and questionnaires are often subjected to aberrant response behaviours (e.g., faking or careless responding). In particular, data affected by faking or careless responding commonly show a deterioration of the inter-item covariance matrix. Based on this evidence, factor mixture models (FMMs) have been recently proposed as screening tools for aberrant responses. FMMs generally include a confirmatory factor analysis (CFA) model, as the unbiased component, and a modified CFA model for the specific type of aberrant response style. In this contribution, we present an FMM using an exploratory factor analysis (EFA) model as the aberrant component (i.e., the CFA+EFA model). By doing so, the CFA+EFA model is a flexible method to detect various types of biased respondents. Indeed, the unconstrained structure of the EFA model captures the biased inter-item correlations that are inconsistent with the target population's latent structure specified in the CFA. Moreover, the CFA+EFA model allows for the inclusion of additional covariates. In this setting, the model estimation is obtained via the Expectation-Maximization algorithm. Finally, we validated our approach by means of simulation and case studies. The results indicate the effectiveness of the proposed FMM.

Latent Variable Models for Panel Data: Comparison of Estimation Methods

Wednesday, 17th July - 09:00: Symposium: Latent variable models for complex data structures (NB A) - Symposia

Dr. Lucia Guastadisegni (University of Bologna), Prof. Silvia Bianconcini (University of Bologna), Prof. Silvia Cagnone (University of Bologna)

In various research fields, latent variable models play a crucial role in offering essential insights into unobservable constructs. These models are particularly valuable when applied to multivariate longitudinal data, as they help elucidate the temporal changes in latent constructs such as attitudes, opinions, performances, and abilities. However, modeling such data poses important challenges, including the discrete nature of observed variables, the impact of item effects, and the need to capture the dynamic evolution of latent constructs over time. One major challenge in estimating latent variable models arises from the computational burden determined by extensive datasets commonly encountered in panel studies with large respondent cohorts, multiple measurement waves, and diverse choice sets. In such situations, traditional likelihood-based and Bayesian estimation approaches become impractical, especially considering the complexities associated with mixed data types, even within cross-sectional models. To address these challenges, this study investigates and compares alternative estimation methodologies, with a specific focus on two promising approaches: composite likelihood methods and dimension-wise quadrature. Composite likelihood methods optimize univariate and/or bivariate likelihood products, reducing the computational demand while maintaining estimation accuracy. Dimension-wise quadrature simplifies high-dimensional integrals by truncating the Taylor series expansion, providing precise approximations without the need for derivative computations. To evaluate and compare the performance of these estimation methods, our research will conduct extensive simulation studies under various empirical scenarios. Additionally, we will demonstrate the practical implications through real data examples, providing insights into the behavior and effectiveness of the proposed estimators.

Scalable M-Estimation for GLLVM

Wednesday, 17th July - 09:00: Symposium: Latent variable models for complex data structures (NB A) - Symposia

Prof. Maria-Pia Victoria-Feser (University of Bologna)

Dimension reduction for high dimensional data is an important and challenging task, relevant to both machine learning and statistical applications. Generalized Linear Latent Variable Models (GLLVMs) provide a probabilistic alternative to matrix factorization when the data are of mixed types, whether discrete, continuous, or a mixture of both. The benefit of GLLVMs is that the model parameters themselves are interpretable and provide meaningful indications on the very structure of the data. Moreover, GLLVM can naturally be extended to dynamic processes such as those used to model longitudinal data. However, with a likelihood-based approach, GLLVM's estimation represents a tremendous challenge for even moderately large dimensions, essentially due to the multiple integrals involved in the likelihood function. Although numerous methods based on approximations of this latter have been proposed, they do not scale well to high dimensions, and they may also introduce a large bias in the estimates. In this paper, we consider an alternative route, which consists in proposing an alternative estimator, based on drastically simplified estimating equations, complemented with a numerically efficient bias reduction methods in order to recover a consistent estimator for the GLLVM parameters. The resulting estimator is an M -estimator, which has a negligible efficiency loss compared to the (exact) MLE. For larger data sets, the proposed M-estimator, whose computational burden is linear in npq, remains applicable when the state-of-the-art method fails to converge. To compute the M -estimator, we propose to use a stochastic approximation algorithm.

Modeling Heterogeneity in Response Strategies

Wednesday, 17th July - 09:00: Symposium: Modeling Heterogeneity in Response Strategies (NB B) - Symposium Overview

Dr. Rudolf Debelak (University of Zurich), Prof. Thorsten Meiser (University of Mannheim)

A central challenge in psychometrics is the accurate modeling of response processes of test takers that lead to observed responses in reaction to a given item. A variety of frameworks are available for modeling such response processes, including item response theory, factor analysis and structural equation modeling, allowing the estimation of model parameters to determine the psychometric characteristics of test takers and items. A central assumption of the model estimations is that the underlying model parameters are stable. A violation of this assumption can lead to serious misinterpretations and has been labeled as “Parameter Heterogeneity” or “Differential Item Functioning” in the literature. In this symposium, we discuss strategies for detecting and addressing such heterogeneity in a variety of different settings and statistical frameworks, encompassing methods from statistics and machine learning. Alagöz and Meiser present the mixture IRTree model as a framework for assigning test takers to classes with unique response mechanisms. Debelak and Meiser develop and apply a family of score-based tests to detect parameter heterogeneity in item response models for response styles. Henninger, Radek, Sengewald and Strobl present a related tree-based approach for detecting violations of parameter invariance in the Partial Credit Model. Kiefer and Mayer finally present SubgroupSEM, a method that combines supervised learning with structural equation modeling that aims to detect and identify groups of test takers who exceptional model parameters.

Mixture IRTree models for exploring heterogeneity in response mechanisms

Wednesday, 17th July - 09:00: Symposium: Modeling Heterogeneity in Response Strategies (NB B) - Symposia

Mr. Ömer Emre Can Alagöz (University of Mannheim), Prof. Thorsten Meiser (University of Mannheim)

IRTree models can improve the validity of our inferences from rating-scale items by accounting for heuristic response strategies such as response styles (RS). Traditional IRTree models decompose ordinal responses into pseudo-items (i.e., nodes) each representing a distinct decision-making process. Each node is then modelled with an item response model. For 4-point items, a response is divided into two nodes: 1) response direction, where the trait affects the probability of agreement, 2) response extremity, where both the trait and extreme RS (ERS) affect the selection of relative (dis)agreement categories.

Despite accounting for RS effects, the traditional models assume that all respondents follow a single response strategy, where the relative (dis)agreement category selections are affected by the trait and ERS to the same extent for all respondents. Because respondents may differ in the extent to which they adopt a heuristic driven strategy (e.g., fatigue, motivation), such assumption of homogeneous response processes is unlikely to be met, which may further result in invalid inferences.

To account for different response mechanisms, we propose the mixture IRTree model (MixTree). In MixTree, participants are assigned to different latent classes, each linked to unique response processes. Given their class memberships, different relative weights are assigned to individuals' trait and ERS scores. Moreover, the MixTree simultaneously analyses extraneous variables to explore the sources of heterogeneity. A simulation study confirmed the performance of the MixTree in recovering classes and model parameters. Empirical data analysis revealed two latent classes, one associated with a trait-driven and the other with RS-driven mechanisms.

Investigating Heterogeneity in IRTree Models for Response Styles Using R

Wednesday, 17th July - 09:00: Symposium: Modeling Heterogeneity in Response Strategies (NB B) - Symposia

Dr. Rudolf Debelak (University of Zurich), Prof. Thorsten Meiser (University of Mannheim)

IRTree Models provide a flexible framework for the modeling of response styles and trait-related judgements in rating scales. Conceptually, IRTree Models consider responses to rating scales as involving multiple processes that can be represented as nodes in a decision tree. Each node is then parametrized by a multidimensional item response theory (IRT) model, representing traits and response styles as latent person parameters. The effectiveness of these models depends on whether the underlying IRT models are valid and that their item parameters are stable over respondent populations. We present a novel approach based on model-based recursive partitioning that aims at detecting and addressing parameter instabilities in IRTree Models by score-based parameter invariance tests. Unlike classical approaches that rely on predefined groups to detect parameter instabilities, our approach allows the flexible detection of parameter instabilities with regard to categorical and continuous person covariates. The detection of such instabilities can be further used to obtain groups of respondents for which the item parameters are stable. In our presentation, we evaluate the new algorithm using simulated and empirical datasets. We further discuss and present an implementation of this algorithm in R, which makes the method widely applicable. The presented method can thus be considered a valuable, accessible tool for researchers and practitioners for investigating response styles in a psychometric framework.

Quantifying DIF and DSF effect sizes in partial credit trees

Wednesday, 17th July - 09:00: Symposium: Modeling Heterogeneity in Response Strategies (NB B) - Symposia

Dr. Mirka Henninger (University of Basel), Jan Radek (University of Basel), Dr. Marie-Ann Sengewald (Leibniz Institute for Educational Trajectories, Bamberg), Prof. Carolin Strobl (University of Zurich)

A modern framework to study differential item functioning (DIF) and differential step functioning (DSF) is model-based recursive partitioning. It combines parametric models, such as the Partial Credit model for polytomous responses, with decision trees from machine learning. The resulting PCtree is a powerful tool to detect previously unknown DIF/DSF groups and allows researchers to use multi-categorical and continuous covariates as background variables. However, the approach cannot identify specific items with DIF/DSF or quantify their effect sizes. To address this limitation, we extend the PCtree method by incorporating an effect size measure for DIF/DSF in polytomous responses. This extension will support researchers in identifying DIF/DSF items, and in evaluating whether a split in a singular PCtree is based on a meaningful difference in item or threshold parameters or could be avoided. Using simulation studies, we assess the effectiveness of the effect size measure in polytomous items under the null hypothesis and in conditions where different forms of DIF/DSF are present. We expect that integrating the effect size measure will prevent unnecessary splits in PCtrees when DIF/DSF effects are negligible, and that it will allow researchers to identify DIF/DSF items. We will also discuss the limitations of the current extensions together with how it can enhance the interpretation of DIF/DSF analyses using PCtrees.

Mining groups with exceptional response processes in structural equation models

Wednesday, 17th July - 09:00: Symposium: Modeling Heterogeneity in Response Strategies (NB B) - Symposia

Christoph Kiefer (Bielefeld University), Axel Mayer (Bielefeld University)

Structural equation modeling (SEM) is one of the most popular statistical frameworks in the social and behavioral science. It is frequently used for the investigation of psychometric scales and/or the examination of their role in a structural model. Often, the detection of groups with distinct sets of parameters in SEM is of key importance for applied researchers – for example, when investigating potential measurement invariance (or: differential item functioning) among individuals for a mental ability test. With the advent of big data, the number of variables potentially accounting for parameter heterogeneity in SEM has increased and, thus, machine learning-based techniques gain in importance for detection of measurement non-invariance. In this talk, we present SubgroupSEM which is a combination of exceptional model mining – a well-established toolkit of supervised learning algorithms and techniques from the field of computer science – with SEM. The main goal of SubgroupSEM is to detect and identify groups with an exceptional set of parameters, for example, an exceptional pattern of intercepts and factor loadings for a latent variable. We provide an introduction to SubgroupSEM and illustrate its distinction from alternative approaches like SEM trees using both artifical and real-world data from a educational large-scale assessment study. We will provide an outlook on recent developements of SubgroupSEM with a focus on applications in psychometric measurement.

Using causal inference theory for designing and analyzing replication studies

Wednesday, 17th July - 09:00: Symposium: Using causal inference theory for designing and analyzing replication studies (NB C) - Symposium Overview

Prof. Steffi Pohl (Freie Universität Berlin), Dr. Marie-Ann Sengewald (Leibniz Institute for Educational Trajectories, Bamberg), Peter M Steiner (University of Maryland), Dr. Vivian Wong (University of Virginia)

So far replication studies may hint at possible reasons for nonreplicability, but the design and analyses of these studies do not allow for a causal claim on which aspects impact effect heterogeneity across studies. In this symposium we present work that applies causal inference theory to replication studies with the aim to evaluate the causal effect of study characteristics (e.g. population, setting, outcome measure) on study results and as such explain effect heterogeneity. In the symposium the causal replication framework, which formalizes the assumptions that need to be fulfilled for an effect to replicate, is presented. Different designs and analyses are presented that allow for identifying the causal effects of study characteristics on study effects, both in uni- and multifactorial designs. The talks in the symposium will present specific designs, show how they can be applied in practice and how assumptions can be tested. Analyses for controlling for unintended study differences as well as analyses on the sensitivity of results are introduced and illustrated in applications. Implications for designing studies are derived and respective analyses tools provided. The research presented in this symposium shows new ways of designing and analyzing replication studies that may help to identify causes of (non-)replicability and as such more clearly specifying the scope and the boundaries of theories.

A causal replication framework for systematic replication efforts

Wednesday, 17th July - 09:00: Symposium: Using causal inference theory for designing and analyzing replication studies (NB C) - Symposia

Peter M Steiner (University of Maryland), Dr. Vivian Wong (University of Virginia)

Despite recent interest to promote the replication of study results, there is not yet consensus on what systematic replication is, how high quality replication studies should be conducted, and which metrics for assessing replication success to choose. This talk addresses these challenges by highlighting methodological considerations for the design and implementation replication studies. We present the Causal Replication Framework (CRF) as a coherent approach for planning and analyzing systematic replications based on two or more studies. Using potential outcomes notation of the Rubin Causal Model, CRF clearly (a) defines causal estimands as the replication target and (b) formalizes the assumptions under which direct replication success can be expected. Direct replication failure occurs when one or more of the causal replication or study-specific assumptions are violated. Given the many and strong assumptions needed for direct replication, it will become clear that successful replications are difficult to achieve in practice. However, an important strength of CRF is that it is straight-forward to prospectively plan different types of conceptual replications that allow researchers to systematically investigate effect homogeneity or heterogeneity across populations, settings and time. CRF also highlights that successful replication efforts strongly depend on theoretical domain knowledge when planning replications. Then, a successful replication does not require that the effect estimates or their testing results replicate but that any observed effect heterogeneities can be causally attributed to the systematically varied factors across studies.

Evaluating coaching supports in teacher education: A replication study approach

Wednesday, 17th July - 09:00: Symposium: Using causal inference theory for designing and analyzing replication studies (NB C) - Symposium Overview

Mrs. Qing Liu (University of Virginia), Dr. Vivian Wong (University of Virginia), Dr. Julie Cohen (University of Virginia)

This study examines the impact of coaching on teacher candidates' pedagogical skills within mixed reality simulation settings. The study design includes two systematic replication efforts designed to evaluate the efficacy and replicability of coaching interventions over systematic sources of variation. In the first study, a "switching replication" design was implemented, where groups alternated between receiving coaching and serving as controls in various simulation scenarios, allowing for an assessment of the robustness of coaching effects across contexts. The second study utilized a multi-site replication approach, investigating the replicability of coaching effects across three distinct teacher preparation with different populations and settings. Combined, findings from these studies aim to contribute insights into the efficacy of coaching for enhancing teaching practices, with an emphasis on understanding the conditions and populations for which these strategies are most beneficial. This presentation will not only delve into the replication design but also discuss the methodological challenges encountered in planning and executing prospective replication studies, alongside strategies for identifying sources of effect heterogeneity. Our work sheds light on the critical question of "what works" in teacher education simulations and underpins the need for rigorous replication to inform educational practices.

A statistical framework for investigating causes of non-replicability

Wednesday, 17th July - 09:00: Symposium: Using causal inference theory for designing and analyzing replication studies (NB C) - Symposia

Prof. Steffi Pohl (Freie Universität Berlin), Dennis Kondzic (Freie Universität Berlin), Jerome Hoffmann (Leibniz Institute for Educational Trajectories, Bamberg), Dr. Marie-Ann Sengewald (Leibniz Institute for Educational Trajectories, Bamberg)

One reason for non-replicability of study results may be that study characteristics, such as population, setting, or outcome measure, vary across primary and replication study. In current replication studies many study characteristics vary at once, making it impossible to infer causes for non-replicability. In this work, we rely on conceptual replications, in which study characteristics are intentionally varied across studies, and focus on identifying the causal effect of study characteristics on the results of a study. In order to estimate the causal effect of study characteristics on the result of a study, one needs to keep everything else but the intentionally varied study characteristic constant across the two studies. Due to practical or ethical reasons, it is not always possible to keep all study characteristics not under investigation constant across studies by design.

In our work we propose a statistical approach that allows for estimating the causal effect of a certain study characteristic on study results even if there are unintended differences between the studies. We show that current approaches for statistically controlling for confounding in group comparisons within a study, cannot necessarily be applied to control for unintended differences in comparisons between studies. We present a) the random probability experiment underlying this process, b) definitions of the effects of interest, c) assumptions needed to identify these effects in empirical studies, and d) the statistical approach for estimating the effect. The approach is illustrated and discussed on an empirical example on replicating an effect in social psychology.

Causes of non-replicability: Examining assumptions for causal inference

Wednesday, 17th July - 09:00: Symposium: Using causal inference theory for designing and analyzing replication studies (NB C) - Symposia

Dr. Marie-Ann Sengewald (Leibniz Institute for Educational Trajectories, Bamberg), Jerome Hoffmann (Leibniz Institute for Educational Trajectories, Bamberg), Dennis Kondzic (Freie Universität Berlin), Dr. Mathias Twardawski (LMU Munich), Dr. Anne Gast (University of Cologne), Johanna Höhs (University of Cologne), Prof. Steffi Pohl (Freie Universität Berlin)

Variations in study characteristics between a primary study and its replication, like differences in effect generating conditions, outcome measures, the target population, or the situation under which a study is conducted, can cause effect heterogeneity. This is formalized in the causal replication framework (CRF) through five assumptions for identifying the same causal effect across different studies. In systematic replications with planned differences between studies, the causal impact on effect heterogeneity can be investigated, provided that all other CRF assumptions are met. Accordingly, we implemented a series of prospective replications and demonstrated how the CRF assumptions can be examined and how unintended differences can be controlled by design or analysis. To achieve this, we clearly define the effect of interest in a treatment-control contrast and intended differences in bivariate study comparisons, including variations in the treatment implementation, the recruited sample, and the time of data collection. Across studies, we kept as many study aspects as possible constant by design. Furthermore, we investigated CRF assumptions by translating approaches for fair group comparisons in primary studies to replication research. We addressed multiple threads for fair comparisons between studies, that is how to (i) control for systematic missing data, (ii) check for balance of treatment stability indicators, (iii) detect differential item functioning in outcome measures, (iv) adjust for unintended differences in person and setting characteristics. Our work demonstrates how recent developments in designing and analyzing replication studies can be applied in practice when the goal is to identify causes of effect heterogeneity.

Fractional factorial replication designs: Implementation and analysis

Wednesday, 17th July - 09:00: Symposium: Using causal inference theory for designing and analyzing replication studies (NB C) - Symposia

Muwon Kwon (University of Maryland), Patrick Sheehan (University of Maryland), Dr. Vivian Wong (University of Virginia), Peter M Steiner (University of Maryland)

When assessing the impact of a new treatment in a randomized experiment or of a potential causal factor in an observational study, a key question often is whether the causal effect replicates and generalizes to different populations and settings. To systematically assess the replicability and generalizability of causal effects we propose the use of fractional factorial replication designs. Such designs rely on all theoretically derived effect moderators to create a factorial design and to estimate the causal effect for each moderator level combination (i.e., for each cell of the design). However, with many moderators, the full factorial design will regularly result in more cells than are feasible to investigate. Under these conditions, researchers can still implement a fractional factorial design, wherein higher-order effects that are assumed to be negligibly small are systematically confounded with main or other higher-order effects. Thus, researchers need to collect data only for a fraction of the entire design. Employing the Federov algorithm in determining a D-optimal fractional design then guarantees that the confounding is minimized while the efficiency of effect estimates is maximized. We demonstrate that fractional factorial designs allow researchers not only to efficiently estimate the causal effects for the observed moderator cells but also to predict the causal effects for cells where no data have been collected. However, the deliberate confounding of higher-order effects requires sensitivity analyses for the predicted effect estimates. Using an example, we demonstrate the implementation and analysis of a fractional factorial replication design.

Advancements in conducting observational intensive longitudinal studies: Design, data, and analysis.

Wednesday, 17th July - 09:00: Symposium: Advancements in conducting observational intensive longitudinal studies: Design, data, and analysis. (NB D) - Symposium Overview

Dr. Sigert Ariens (KU Leuven)

Spurred by increasing interest in the study of intraindividual variability in psychological processes over time, designs which lead to the collection of intensive longitudinal data (ILD) have become popular in many fields of psychological inquiry. Observational designs such ambulatory assessment, experience sampling, and daily diaries are often employed to gather information about how psychological processes evolve over time. In this symposium, we highlight recent methodological developments with important implications for the design of such studies and analysis of these data. Revol, J. will set the stage with a presentation of a preprocessing pipeline, offering researchers a transparent method for checking the adequacy of their data and informing adequate preprocessing steps for analysis. Ariens, S. will proceed with a talk on state of the art designs aimed at optimally estimating autoregressive dynamics in the data. In a similar vein, Simsa, B. will present a study on the implications of different missing data mechanisms for estimating the fixed autoregressive effect in multilevel contexts. Leertouwer, I. continues the topic of design with a presentation on different methods for studying the reliability of ambulatory assessment data. Haqiqatkhah, M. closes with an investigation on day-of-week effects and seasonal dynamics in daily diary data by proposing a framework for understanding them with visualizations and modeling

Preprocessing ESM data: A framework, tutorial website, R package, and reporting templates

Wednesday, 17th July - 09:00: Symposium: Advancements in conducting observational intensive longitudinal studies: Design, data, and analysis. (NB D) - Symposia

Jordan Revol (KU Leuven), Chiara Carlier (KU Leuven), Prof. Ginette Lafit (KU Leuven), Dr. Martine Verhees (KU Leuven), Dr. Laura Sels (Ghent University), Prof. Eva Ceulemans (KU Leuven)

Experience Sampling Method (ESM) studies have become a very popular tool to gain insight into the dynamics of psychological processes. Whereas the statistical modeling of ESM data has been widely studied, the preprocessing steps that precede such modeling have received relatively limited attention, despite being a challenging phase. At the same time, adequate preprocessing of ESM data is crucial: it provides valuable information about the quality of the data and, importantly, helps to resolve issues in the data that may compromise the validity of statistical analyses. To support researchers in properly preprocessing ESM data, we have developed a step-by-step framework, a tutorial website that provides a gallery of R code, an R package, and templates to report the preprocessing steps. Particular attention is given to three different aspects in preprocessing: checking adherence to the study design (e.g., were the momentary questionnaires delivered according to the sampling scheme), examining participants' response behaviors (e.g., compliance, careless responding), and describing and visualizing the data (e.g., examining distributions of variables).

Episode-contingent experience-sampling designs for accurate estimates of autoregressive dynamics

Wednesday, 17th July - 09:00: Symposium: Advancements in conducting observational intensive longitudinal studies: Design, data, and analysis. (NB D) - Symposia

Dr. Sigert Ariens (KU Leuven), Jordan Revol (KU Leuven), Prof. Ginette Lafit (KU Leuven), Dr. Janne Adolf (KU Leuven), Prof. Eva Ceulemans (KU Leuven)

Affect dynamics are often studied by means of first-order autoregressive (AR) modeling applied to intensive longitudinal data. A key target in these studies is the AR parameter, which is often tied conceptually to regulatory behavior in the affective process. The data is typically gathered using experience sampling methods, which are designed to pick up on fluctuations in affective variables as they evolve over time in naturalistic settings. In this talk, we present a manuscript where we compared classical time-contingent sampling designs to episode-contingent sampling designs, which initiate sampling when an emotional episode has been signaled. We define emotional episodes as periods where an affective process strays relatively far away from its mean. Compared to time-contingent designs, episode-contingent designs leverage on increased affective variability, which can have beneficial implications for the precision of the ordinary least squares AR effect estimator. Using an extensive simulation study, we attempt to delineate which characteristics of an episode-contingent design are important to consider, and how these characteristics are related to estimation benefits. We conclude that episode-contingent designs can have marked benefits for the precision of the AR effect estimator, and discuss a number of challenges when it comes to implementing episode-contingent designs in practice.

Consequences of missing data types of the multilevel AR(1) model

Wednesday, 17th July - 09:00: Symposium: Advancements in conducting observational intensive longitudinal studies: Design, data, and analysis. (NB D) - Symposia

Mr. Benjamin Simsa (Charles University), Jordan Revol (KU Leuven), Prof. Ginette Lafit (KU Leuven), Ms. Leonie Cloos (KU Leuven), Dr. Janne Adolf (KU Leuven), Prof. Eva Ceulemans (KU Leuven)

Missing data is a challenging issue in many study designs. This issue is notably prevalent in the Experience Sampling Method (ESM), where missed observations are generally expected to occur and manifest in diverse ways. We investigated the effect of specific missingness mechanisms (i.e., missing completely at random versus tail-based missingness), temporal patterns (whether or not missings occur consecutively), and compliance level on the bias and variance of a commonly used model in ESM studies, the multilevel autoregressive (MLAR) model. In a simulation study, we showed that bias and variance due to missing data are more severe when compliance is low and when the missingness is tail-based (ie, when the missingness depends on the process value itself). We corroborate these results by setting data of an empirical ESM dataset to missing following the same principles. To lower the estimation bias and variance, we particularly recommend making design choices to promote compliance and/or increase the probability of data being missing completely at random: decreasing participant burden by using shorter questionnaires, providing adequate incentives to participants, and considering using planned missingness designs or episode-based sampling.

Combining day-to-day dynamics with day-of-week effects and weekly dynamics using seasonal ARMA models

Wednesday, 17th July - 09:00: Symposium: Advancements in conducting observational intensive longitudinal studies: Design, data, and analysis. (NB D) - Symposia

Mr. Mohammadhossein Manuel Haqiqatkhah (Utrecht University), Prof. Ellen L. Hamaker (Utrecht University)

Daily diary data of emotional experiences are typically modeled with a first-order autoregressive model to account for possible day-to-day dynamics. However, our emotional experiences are likely to be affected by the weekly rhythm of our activities, which may be reflected by: (a) day-of-week effects (DOWEs), where different days of the week are characterized by different means; and (b) week-to-week dynamics, where weekday-specific activities and experiences have a delayed effect on the emotions that we experience on the same weekday a week later. While DOWEs have been studied occasionally, week-to-week dynamics have been largely ignored in psychological research. To gain more insight in the various regularities that may exist in daily diary data, we begin with presenting a set of complementary visualization techniques that can help to detect and characterize weekly rhythms and day-to-day dynamics in time series data. Subsequently, we introduce the family of seasonal autoregressive–moving average (SARMA) models from the econometrics literature, and extend this with models for the DOWEs. We illustrate how the different model components show up in the various visualizations of the time series data. We then provide a tutorial on fitting these models in R, discussing model fit and model selection, and apply this to a daily diary dataset consisting of 56-101 daily measures from 98 individuals. The results suggests that most individuals in the sample are characterized by patterns and dynamics that the current practices in psychological research cannot capture adequately. We discuss the implications of our findings for current psychological research practices.

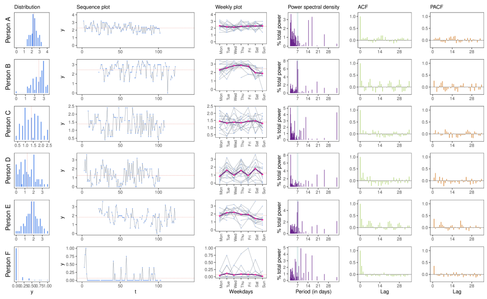


Figure 1

Marginal distribution, sequence, weekly, power spectral density (PSD), autocorrelation function (ACF), and partial autocorrelation function (PACF) plots for the empirical data of six individuals in Wright et al. (2015). Note that the time index in the sequence plots corresponds to the days since the start of the study.

Fig 1.png

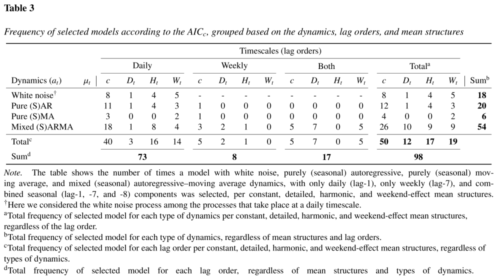


Table 3

Frequency of selected models according to the AIC$_c$, grouped based on the dynamics, lag orders, and mean structures

	Timescales (lag orders)																
	Daily				Weekly				Both				Total[a]				
Dynamics (a_t) μ_t	c	D_t	H_t	W_t	c	D_t	H_t	W_t	c	D_t	H_t	W_t	c	D_t	H_t	W_t	Sum[b]
White noise[†]	8	1	4	5	-	-	-	-	-	-	-	-	8	1	4	5	**18**
Pure (S)AR	11	1	4	3	1	0	0	0	0	0	0	0	12	1	4	3	**20**
Pure (S)MA	3	0	0	2	1	0	0	0	0	0	0	0	4	0	0	2	**6**
Mixed (S)ARMA	18	1	8	4	3	2	1	0	5	7	0	5	26	10	9	9	**54**
Total[c]	40	3	16	14	5	2	1	0	5	7	0	5	**50**	**12**	**17**	**19**	
Sum[d]	**73**				**8**				**17**				**98**				

Note. The table shows the number of times a model with white noise, purely (seasonal) autoregressive, purely (seasonal) moving average, and mixed (seasonal) autoregressive–moving average dynamics, with only daily (lag-1), only weekly (lag-7), and combined seasonal (lag-1, -7, and -8) components was selected, per constant, detailed, harmonic, and weekend-effect mean structures.
[†] Here we considered the white noise process among the processes that take place at a daily timescale.
[a] Total frequency of selected model for each type of dynamics per constant, detailed, harmonic, and weekend-effect mean structures, regardless of the lag order.
[b] Total frequency of selected model for each type of dynamics, regardless of mean structures and lag orders.
[c] Total frequency of selected model for each lag order per constant, detailed, harmonic, and weekend-effect mean structures, regardless of types of dynamics.
[d] Total frequency of selected model for each lag order, regardless of mean structures and types of dynamics.

Tab 3.png

A practical guide to estimating the reliability of ambulatory assessment data

Wednesday, 17th July - 09:00: Symposium: Advancements in conducting observational intensive longitudinal studies: Design, data, and analysis. (NB D) - Symposia

Dr. IJsbrand Leertouwer (Erasmus University Rotterdam)

Spurred by increasing interest in the study of intraindividual variability in psychological processes over time, designs which lead to the collection of intensive longitudinal data (ILD) have become popular in many fields of psychological inquiry. Observational designs such ambulatory assessment and experience sampling are often employed to gather information about how psychological processes evolve over time. In this symposium, we highlight recent methodological developments with important implications for the design of such studies and analysis of these data. Revol, J. will set the stage with a presentation of a preprocessing pipeline, offering researchers a transparent method for checking the adequacy of their data and informing adequate preprocessing steps for analysis. Ariens, S. will proceed with a talk on state of the art designs aimed at optimally estimating autoregressive dynamics in the data. In a similar vein, Simsa, B. will present a study on the implications of different missing data mechanisms for estimating the fixed autoregressive effect in multilevel contexts. Leertouwer, I. continues the topic of design with a presentation on different methods for studying the reliability of ambulatory assessment data. Haqiqatkhah, M. closes with an investigation on seasonal dynamics in ambulatory assessment data by applying appropriate analytic methods.

Quantitative methods for assessing the impact of physical environment on health

Wednesday, 17th July - 09:00: Symposium: Quantitative methods for assessing the impact of physical environment on health (RB 209) - Symposium Overview

Dr. Sjoerd Ebisch (università degli)

This symposium addresses new developments in psychometric and analytical methods to investigate how aspects of the physical environment influence mental health and wellbeing. Multidisciplinary studies from environmental psychology, epidemiology, ecology, public and environmental health sciences consistently report beneficial effects of natural (es., green or blue) spaces on health. By contrast, urban (es., grey) spaces can have a negative impact. This growing field of research is of considerable scientific as well as public interest but faces a series of methodological challenges to obtain critical insights with increasing detail and applicability. Specifically, the current state of research requires a progressive quantification of general to specific environmental features facilitating practical implementation, to calibrate the relationship between objective environmental features and subjective perception, to quantify the perceived environmental features in different facets of everyday life (e.g., home, work, study, travel, spare time), to elucidate the psychological mechanisms involved in the relationship between environmental and health variables and their contribution to the risk of psychopathological phenomena, as well as to identify individual predispositions that influence the environmental impact on health. The contributions to this symposium provide complementary insights in these topics, by presenting an innovative psychometric instrument to assess restorative qualities of physical environments in work contexts, the implementation of environmental experiences to ease athletes' recovery comparing psychological and physiological measures, and the development and application of sophisticated environmental image clustering methods together with subjective perception to predict mental health.

Predicting Mental Health Trough Satellite Images

Wednesday, 17th July - 09:00: Symposium: Quantitative methods for assessing the impact of physical environment on health (RB 209) - Symposia

Dr. Simone Di Plinio (G. d'Annunzio University of Chieti and Pescara), Dr. Elisa Menardo (University of Verona), Dr. Daniela Cardone (G. d'Annunzio University of Chieti and Pescara), Dr. Claudia Greco (G. d'Annunzio University of Chieti and Pescara), Prof. Arcangelo Merla (G. d'Annunzio University of Chieti and Pescara), Prof. Margherita Brondino (University of Verona), Prof. Margherita Pasini (University of Verona), Prof. Sjoerd Ebisch (G. d'Annunzio University of Chieti and Pescara)

Empirical studies indicate that natural landscapes, such as coastlines and forests, provide greater restorative benefits compared to urban environments. Factors like individual identity can influence the level of perceived restorativeness. Our research aimed to establish standardized, straightforward procedures for assessing the impact of both objective and subjective elements on human health. We investigated how natural versus urban settings affect personal restorative experiences, engaging around 1,000 participants from Italy. Utilizing both psychometric evaluations and analytical approaches, we examined the influence of the physical characteristics of participants' local environments on their perceived restorativeness. We developed an original pipeline for the analysis of satellite imagery, through which we quantitatively assessed the greenery and urban elements surrounding the participants' residences. Our sophisticated analysis included a comprehensive image analysis framework and iterative clustering techniques for categorizing environmental features. Our results, confirmed by multivariate analysis, reveal that green spaces significantly enhance restorativeness, in stark contrast to urban spaces, which negatively impact it. The positive effects of green spaces were especially pronounced in promoting feelings of Fascination, Being-Away, and Scope, with these benefits being influenced by the individual's sense of identity. This study not only highlights the critical role of environmental characteristics in promoting well-being but also introduces novel methodologies for forecasting the well-being impacts of environmental changes.

Assessing recovery after high-intensity anaerobic exercise: a comparison between physiological and psychological measures

Wednesday, 17th July - 09:00: Symposium: Quantitative methods for assessing the impact of physical environment on health (RB 209) - Symposia

Dr. Luca Laezza (University of Verona), Dr. Carlos de la Torre Perez (Universidad Autónoma de Madrid), Prof. Victor Rubio (Universidad Autónoma de Madrid), Prof. Stefano De Dominicis (University of Copenhagen), Dr. Martina Vacondio (University of Verona), Dr. Alessandro Fornasiero (University of Verona), Prof. Barbara Pellegrini (University of Verona), Prof. Margherita Brondino (University of Verona)

Drawing from established theories in environmental psychology, including Stress Reduction Theory and Attention Restoration Theory, which underscore the restorative potential of natural environments, we seek to understand the effects of restorative environments on athletic performance and health.

This study will assess the reliability of psychological measures compared to physiological measures, examining the influence of exposure to restorative natural environments on athlete recovery following anaerobic exercise through a mixed design. Data collection is still ongoing in three different countries: Italy, Spain and Denmark.

Participants engage in anaerobic exercise on stationary bikes and are then exposed to a 3-minute video showing either a restorative natural environment or an urban non-restorative environment. The restorativeness of both videos was previously assessed on a different sample. Pre and post-exercise measures encompass muscular, metabolic, and autonomic indices to comprehensively assess physiological responses. Psychological measures include core affect, perceived effort, state anxiety, psychological restoration and perceived restorativeness of the environment.

Through rigorous experimental design and meticulous data collection, this study aims to elucidate whether exposure to restorative natural environments yields discernible impacts on physiological markers of recovery in athletes.

Findings from this investigation will contribute to clarifying the effectiveness of psychological and physiological measures within the framework of studying the restorativeness benefits of natural environments on wellbeing and mental health.

Furthermore, novel insights into the potential physiological benefits of restorative environments within the context of athletic recovery offer valuable implications for sports science and performance optimisation strategies.

Validation of the “Restorativeness at Work Scale” (R@WS)

Wednesday, 17th July - 09:00: Symposium: Quantitative methods for assessing the impact of physical environment on health (RB 209) - Symposia

Prof. Margherita Brondino (University of Verona), Dr. Elisa Menardo (University of Verona), Dr. Camilla Marossi (University of Verona), Prof. Margherita Pasini (University of Verona)

This research aims to build a scale, called Restorativeness at work scale (R@WS), to measure the perception of the restorative qualities of physical environments in work contexts, and to study its psychometric properties. Physical environments can affect the individual’s ability to direct attention by consuming them, but “restorative environments” can relieve mental fatigue.

This work was carried out through different steps using a mixed methodology (qualitative and quantitative methods). First, 20 semi-structured interviews were carried out and administered to two types of workers: office and production. The questions aimed to investigate and deepen the relevant elements in the workplace for the construct of restorativeness. Second, we moved on to constructing the items, taking the PRS (PRS, Pasini et al., 2014) items as a reference and taking inspiration from the themes that emerged in the first step. 27 potential items were identified to investigate 4 dimensions (FA, B-A, COH, SCO); it has been chosen to exclude Compatibility. The first version of the scale with 27 items was administered to 28 workers. Each participant also underwent a cognitive interview aimed at identifying any critical issues. Based on the previous step’s results, we refined or excluded problematic items. A second version of the scale (17 items) was administered to 673 workers (61% female), and 238 of these (61% female) completed the scale a second time after about one month. Psychometrics proprieties (factor structure, internal consistency, test-retest reliability) were verified on this data.

Comparing EIRT models for Mathematics and ELA across grades 3-8

Wednesday, 17th July - 09:00: Applications of IRT (RB 210) - Oral

Dr. Magdalen Beiting-Parrish (Federation of American Scientists), Dr. Sydne McCluskey (private researcher), Dr. Jay Verkuilen (City University of New York), Dr. Howard Everson (City University of New York)

Standardized tests frequently determine students' educational futures, yet for many of these exams, students from demographic minority backgrounds receive lower scores, on average, compared to their majority background peers. One potential source of this may be the reading comprehension component of these exams, especially when reading comprehension is not the primary skill being tested (Messick, 1989). Previous research has shown that changing linguistic characteristics of tests can improve test performance for demographic minority groups; for example, Abedi et al. (2003, 2000, 1998) found that decreasing item wordiness increased English Language Learner performance by as much as 49%. Additionally, other research has found that other linguistic features such as: polysemous words (Martiniello, 2009), increased use of abstract and academic concept words (Molina, 2012) and complex grammatical structures (Nagy & Townsend, 2012; Shaftel et al., 2006), all contribute to decreased student comprehension and performance for items with these components. This study uses data from a large-scale Northeastern testing program for grades 3-8 for both Mathematics and English Language Arts to better understand the relationship between the linguistic features of the test items, student demographic characteristics, and student performance using Explanatory Item Response models. The primary research goal is to investigate how linguistic features of test items differentially impact student performance across grades, within content standards/sub-domains, and between the two content areas. This study aims to better understand how the language of test items from different content areas impacts student performance to establish quantitative justifications for considering linguistic equity and accessibility of items.

Using item response theory to investigate whether rater assessments measure rater quality: Is there such a thing as a "correct" rating?

Wednesday, 17th July - 09:15: Applications of IRT (RB 210) - Oral

Dr. William Belzak (Duolingo), Dr. Yigal Attali (Duolingo), Ms. Dani Mann (Duolingo)

Rating tasks are common in the behavioral and social sciences, and across many industries. For example, the development of ChatGPT relied on humans to evaluate the quality of output text and assign scores according to a standard rubric; these scores were then used to fine-tune model output. As part of large-scale rating tasks, assessments are used to evaluate rater quality and ensure that raters are making appropriate ratings. However, evaluating rater quality through assessment implies that there are "correct" answers to rating tasks, where "correctness" is typically derived from an "expert" (or from a consensus of experts). In this talk, we investigate whether certain types of rating tasks have "correct" answers and whether rater assessments measure what they intend to measure (i.e., rater quality). Our application involves remote proctors (that serve as raters) of a high-stakes English proficiency test who make decisions about whether test takers have violated rules during test sessions (that serve as rating tasks). We use item response theory (IRT) to demonstrate that defining "correctness" is not always defensible for certain types of rating tasks. Namely, divergent estimates of item discrimination and low estimates of test-retest reliability suggest that some rater assessments fail to measure rater quality. Alternative standards with which to judge rater quality, such as rating "severity" (determined by raters), rating "reasonableness" (determined by experts), or rating in "agreement with policy" (determined by policy-makers), may prove more tractable as measurement goals.

Selection of Eye Tracking Stimuli via Item Response Theory

Wednesday, 17th July - 09:30: Applications of IRT (RB 210) - Oral

Benjamin Graves (University of Missouri), Prof. Edgar Merkle (University of Missouri)

Eye tracking studies often consist of a large pool of images being presented to a participant over multiple experimental sessions. This takes up valuable time and resources from both the participants and researcher. The goal of this project is to explore the application of item response models to eye tracking data in order to assess whether the number of images used in a study can be reduced and still provide similar amounts of information. Specifically, I use data from an image memorability study by Bylinskii et al. (2015) to fit item response models formulated in a GLMM framework. Associated memorability scores are used as a standard of comparison for parameter estimates in the item response models. This method provides a way to select images of varying difficulty and to thin out images that overlap in the information they provide. Alternative link functions are explored for use with eye tracking data that is not binary and simulations are conducted to assess various thinning methods as well as their stability. Overall, models tend to retain their predictive ability as the number of images are reduced. These findings suggest that researchers can decrease the number of images used in a study, given that they are high quality and cover a range of difficulty levels. This decrease in images then saves the resources of both the researcher and participant.

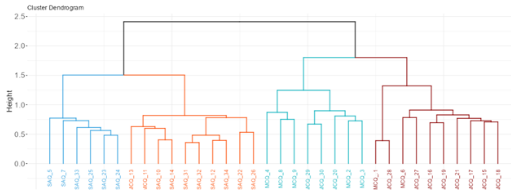
Modernizing High School Assessments in Albania: Exploring Item Response Theory in State Matura Exams

Wednesday, 17th July - 09:45: Applications of IRT (RB 210) - Oral

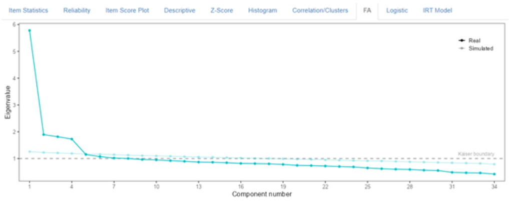
Prof. Eralda Gjika (University of Tirana), Dr. João Paulo Araujo Lessa (University Tuiuti do Paraná), Dr. Afërdita Alizoti (Centre of Educational Services), Dr. Lule Basha (University of Tirana)

In Albania, high school students undergo testing upon completion of their three-year studies, with results determining university entrance based on a meritocratic system. The State Matura exams, administered in a traditional pencil and paper format since 2006, encompass three mandatory exams: Foreign language (English, French, German, Italian, Spanish, and Turkish), Albanian language and literature, and Mathematics, along with one elective exam from a list of eight. Starting from 2019, all tests contain a total of 60 scores, with 20 scores designated for multiple-choice questions and 40 scores for open response questions, including structured and essay-type questions in Albanian literacy and foreign language. These items have low to high difficulty levels, estimated mostly on exam developers' experience. So far, Classical Test Theory (CTT) approach is employed to estimate student proficiency levels, using a decimal grading system ranging from 4 to 10. This study investigates the feasibility and limitations of implementing Item Response Theory (IRT) models in Albanian high school examinations. Our study focuses on one of the elective exams, recognizing the limitation of small sample size, the absence of historical item parameters, and the shortage of trained specialists in the IRT approach.

The expected results include insights into the effectiveness of IRT models in improving the accuracy of student proficiency assessment, identification of potential challenges in calibration and application, and preliminary evidence supporting the modernization of Albania's e-assessment framework. A ShinyApp designed for the corresponding methodology will be presented for the public.



Cluster items.png



Fa.png

Enhancing Scoring Procedures for the Divergent Association Task: Leveraging Sequential Scoring and IRT Paradigm

Wednesday, 17th July - 10:00: Applications of IRT (RB 210) - Oral

Dr. Daniel Dostál (Department of Psychology, Faculty of Arts, Palacký University Olomouc), Kryštof Petr (Department of Psychology, Faculty of Arts, Palacký University Olomouc)

The Divergent Association Task (DAT) stands as a concise computer-administered creativity assessment tool (Olson et al., 2021), prompting participants to devise 10 nouns showcasing maximal dissimilarity across all conceivable meanings and applications. Its hallmark lies in its capacity for objective computerized scoring, wherein each response is mapped onto a multidimensional semantic space via word embeddings (GloVe algorithm). The method's authors advocate for evaluating test performance through the average cosine distance between all pairs of provided words.

This paper critically examines the existing scoring methodology and proposes several alternative approaches grounded in sequential scoring of responses. These novel procedures retain the test's inherent advantages while embracing the benefits inherent in transitioning to the Item Response Theory (IRT) paradigm. Our analysis encompasses a diverse sample of over 1,700 Czech participants who were administered a range of creativity tests, including the DAT. Through empirical demonstrations, we discuss the effectiveness of each proposed assessment procedure, contributing to the ongoing debate on optimizing creativity assessment methodologies.

A Bayesian Approach for Joint Modeling of Rankings and Ratings

Wednesday, 17th July - 09:00: Bayesian Issues in IRT (RB 211) - Oral

Dr. Michael Pearce (Reed College), Prof. Elena Erosheva (University of Washington)

Rankings and ratings are commonly used to express preferences but provide distinct and complementary information. Rankings give ordinal and scale-free comparisons but lack granularity; ratings provide cardinal and granular assessments but may be highly subjective or inconsistent. Collecting and analyzing rankings and ratings jointly has not been performed until recently due to a lack of principled methods. In this work, we propose a flexible, joint statistical model for rankings and ratings under heterogeneous preferences: the Bradley-Terry-Luce Binomial (BTL-Binomial). We employ a Bayesian mixture of finite mixtures (MFM) approach to estimate heterogeneous preferences, understand their inherent uncertainty, and make accurate decisions based on ranking and ratings jointly. We demonstrate the efficiency and practicality of the BTL-Binomial MFM approach on real and simulated datasets of ranking and rating preferences in peer review and survey data contexts.

Bayesian nonparametric HETOP models for rating data

Wednesday, 17th July - 09:15: Bayesian Issues in IRT (RB 211) - Oral

Dr. Giuseppe Mignemi (Bocconi Institute for Data Science and Analytics)

Rating procedure is crucial in many applied fields (e.g., educational, clinical, emergency). It implies that a rater (e.g., teacher, doctor) rates a subject (e.g., student, doctor) on a rating scale. Given raters' variability, several statistical methods have been proposed for assessing and improving the quality of ratings (Gwet, 2014). The analysis and the estimate of inter-rater reliability (IRR) are major concerns in such cases. As evidenced by Martinkova et al. (2023), inter-rater reliability might differ across different subgroups of raters and might be affected by contextual factors. Estimating IRR in the presence of heterogeneity has been one of the recent challenges in this research line. Consequently, several models have been proposed to address this issue under a parametric multilevel modelling framework, in which strong distributional assumptions are made. We propose a more flexible model under the Bayesian nonparametric (BNP) framework, in which most of those assumptions are relaxed. Through hierarchical discrete nonparametric priors, the model accommodates clusters among both raters and subjects and naturally accounts for heterogeneity. Focusing on the general case in which ratings are on an ordinal scale, we propose a BNP heteroscedastic ordered probit (HETOP) model which allows us to estimate different types of IRR indexes and an inter-rater polarization index. The latter gives additional information concerning the polarization of raters' systematic biases. A real-world application is presented and possible future directions are discussed.

Some posterior standard deviations of the graded response model

Wednesday, 17th July - 09:30: Bayesian Issues in IRT (RB 211) - Oral

Prof. Seock-Ho Kim (University of Georgia)

The procedures required to obtain the approximate posterior standard deviations of the parameters of the graded response model in item response theory for polytomous items are described and used to generate values for some common situations. The results are compared with those obtained from maximum likelihood estimation. It is shown that the use of priors may reduce the instability of estimates of the item parameters assuming that the choice of priors is reasonable. The sample size required for acceptable accuracy for the purposes of practical applications of the graded response model may be inferred from tables or computer programs. It is suggested that the careful selection of priors be exercised to obtain the required precision when the graded response model is employed in practical applications.

Estimation of IRT models using NUTS under non-normal distributions

Wednesday, 17th July - 09:45: Bayesian Issues in IRT (RB 211) - Oral

Dr. Rehab AlHakmani (Emirates College for Advanced Education), Prof. Yanyan Sheng (The University of Chicago)

IRT models are typically estimated by assuming a normal distribution for the person latent trait parameter(s) in the population, which is often plausible given the sample size or latent traits under study in educational and psychological measurement. However, for measuring health, personality, or psychopathology constructs when sample sizes are relatively small, it is unreasonable to assume normally distributed latent traits. Prior research noted bias in estimating IRT models using maximum likelihood estimation when normality is violated (e.g., Zwinderman & Wollenberg, 1990). It is, however, not clear whether similar results apply to the fully Bayesian estimation, especially with the use of non-random walk Markov chain Monte Carlo (MCMC) algorithms. This study hence focuses on evaluating the performance of such algorithm, namely, the no-U-turn sampler (NUTS), in recovering item parameters of the two-parameter logistic (2PL) model under non-normality. Monte Carlo simulations were carried out for test situations with varying sample sizes, test lengths, prior specifications, and degrees of non-normality. Non-normal distributions were selected to be comparable to those adopted by others (e.g., Le & Adams, 2013) and were generated using a third-order polynomial transformation method (Fleishman, 1978). With 25 replications, the accuracy of parameter recovery was evaluated using relative bias and relative root mean square error. Results of the study provide researchers and practitioners with guidelines on the performance of NUTS in estimating a conventional IRT model under situations where trait distributions deviate from normality. They further shed light on how sample sizes, test lengths, and/or prior specifications help mitigate the problem.

A Bayesian model to represent ambiguity and atypical behaviour in Item Response Theory

Wednesday, 17th July - 10:00: Bayesian Issues in IRT (RB 211) - Oral

Mario Angelelli (University of Salento), Enrico Ciavolino (University of Salento), Serena Arima (University of Salento)

Model uncertainty is a fundamental problem in psychometry when multiple explanatory models obtained from observed variables coexist. This represents an indeterminacy factor, which undermines the identification of latent traits that represent conceptual constructs.

This contribution introduces a new methodological framework to investigate ambiguity in psychometric measurements with ordinal responses. The proposal is based on an information-theoretic model that combines multiple probability distributions to describe ambiguity in terms of deviations from rational behavior, as formalized by Luce's axioms. We specify this model by generalizing Ellsberg's paradox, which is a paradigm of decision-making under ambiguity, and examine model symmetries (translations and permutations) as a source of model indeterminacy.

These premises are exploited to represent ambiguity in Item Response Theory (IRT) and Graded Response Models (GRM). Specifically, we propose a Bayesian hierarchical model to specify the latent traits that explain individuals' responses while taking into account inconsistent ordering in the items, in particular differences in the perception of items' difficulty. To enhance identifiability and detection of atypical behaviors, we explore the use of global-local priors for latent traits. The proposal is analyzed through an extensive set of simulations, and we discuss its applicability to capability assessment, focusing on big data-driven organizations. In this context, atypical behaviors highlighted in maturity models could be relevant in the interaction with emerging technologies and potential sources of innovation.

A Selective Intellectual History of Differential Item Functioning Analysis in Item Response Theory and Factorial Invariance in Factor Analysis

Wednesday, 17th July - 10:30: Career Lifetime Achievement Award (Vencovského aula) - Career Lifetime Achievement Award

Prof. David Thissen (University of North Carolina at Chapel Hill)

The concept of factorial invariance has evolved since it originated in the 1930s as a criterion for the usefulness of the multiple factor model; it has become a form of analysis supporting the validity of inferences about group differences on underlying latent variables. The analysis of differential item functioning (DIF) arose in the literature of item response theory (IRT), where its original purpose was the detection and removal of test items that are differentially difficult for, or biased against, one subpopulation or another. The two traditions merge at the level of the underlying latent variable model, but their separate origins and different purposes have led them to differ in details of terminology and procedure. This review traces some aspects of the histories of the two traditions, ultimately drawing some conclusions about how analysts may draw on elements of both, and how the nature of the research question determines the procedures used. Whether statistical tests are grouped by parameter (as in studies of factorial invariance) or across parameters by variable (as in DIF analysis) depends on the context and is independent of the model, as are subtle aspects of the order of the tests. In any case in which DIF or partial invariance is a possibility, the invariant parameters, or anchor items in DIF analysis, are best selected in an interplay between the statistics and judgment about what is being measured.

Paths to the future: A panel discussion of the future of SEM

Wednesday, 17th July - 13:00: Invited Panel: Paths to the future: A panel discussion of the future of SEM (Vencovského aula) - Invited Panel

Prof. Steven Boker (University of Virginia), Prof. Michael Neale (Virginia Commonwealth University), Prof. Yves Rosseel (Ghent University), Prof. Timo von Oertzen (Max Planck Institute for Human Development Berlin), Prof. David Kaplan (University of Wisconsin - Madison), Prof. Timothy Brick (The Pennsylvania State University), Dr. Michael Hunter (The Pennsylvania State University), Prof. Alberto Maydeu-Olivares (University of South Carolina)

This is a panel discussion brings together authors of three of the popular open source SEM programs and other experts who have pushed the boundaries of SEM methods in order to explore how SEM may evolve. Ideas such as “operator nodes”, incorporating neural network models, varieties of regularization procedures, Bayesian estimation, new optimization algorithms, products of variables for nonlinear SEM, and incorporation of generative large language models similar to GPT will be open for discussion. We hope to have a lively interaction between members of the panel and audience.

Individual Salience Weights in the Multidimenisonal Generalized Graded Unfolding Model

Wednesday, 17th July - 13:00: Topics in IRT 1 (NB A) - Oral

Prof. James Roberts (Georgia Institute of Technology)

The multidimensional generalized graded unfolding model (MGGUM) is an item response theory model for analyzing preference-like or self-similarity responses to stimuli with complex multidimensional structure. Like classical multidimensional unfolding models such as the weighted Euclidean model (WEM), the MGGUM produces a configuration in which both respondents and stimuli are simultaneously located in a multidimensional latent space such that proximity between a respondent and stimulus suggests a higher expected response. However, the MGGUM and WEM differ in meaningful ways. First, estimates of MGGUM parameters are based on averages or modes from marginal or fully Bayesian posterior distributions, whereas WEM parameters are estimated by minimizing a badness-of-fit criterion. Second, the MGGUM uses binary or graded ratings of preference for or self-similarity to stimuli. In contrast, the WEM works best with rankings of these same constructs. (Ratings can work well, but ties may become an issue.) Finally, the MGGUM has dimension weights that vary for each item (i.e., discrimination parameters) but the WEM has dimension weights that vary for each respondent (i.e., salience parameters). This presentation focuses on the latter difference. To make these models more comparable across classical and IRT domains, a hybrid MGGUM is developed using person salience parameters rather than item discrimination parameters. The estimates of person location, item location and person salience parameters from the hybrid model are compared with those from an analogous WEM. This comparison is conducted using simulated data and also a large, real data set that contains physical attractiveness ratings for computer generated, human-like stimuli.

Item response function variability: A strategy for model comparison research

Wednesday, 17th July - 13:15: Topics in IRT 1 (NB A) - Oral

Mr. Xing Chen (Fordham University), Dr. Leah Feuerstahler (Fordham University)

Simulation studies are a commonly used method in IRT model comparison research. However, there are few established guidelines on how to generate data from different models in ways that do not introduce confounds between the different models. Although data generated from different models may appear to be drawn from similar distributions, there may be systematic differences among the generated curves. For example, using the same data-generating distributions of discriminations (a) and difficulties (b) for the two- and three-parameter models (2PL and 3PL) will lead to systematically steeper curves for the 2PL. In this paper, we introduce item response function (IRF) variability as a way to simulate data from different IRT models in a way that avoids unnecessary systematic differences between models. Seven symmetric and asymmetric models were investigated in this research. The variation was quantified using distributions of the highest slope and inflection point location implied by each model. In the simulation study, we simulated data using two different methods: first, simulate data keeping discriminations (a) and difficulties (b) constant; second, simulate data keeping IRF variation as similar as possible by adjusting the data-generating parameters. By comparing the accuracy of the simulation study results, it is possible to infer the parameterizing effect of different models when simulating data. In real data analysis, investigating IRF variability through estimated parameters may be useful to better contextualize the model comparison results. The suggested procedure is illustrated with TIMSS data.

Refinement of the Matching Item Response Model: A Shift to Generative Modeling

Wednesday, 17th July - 13:30: Topics in IRT 1 (NB A) - Oral

Mr. Kentaro Fukushima (The University of Tokyo), Ms. Rinhi Higashiguchi (The University of Tokyo), Dr. Kensuke Okada (The University of Tokyo)

Matching format test items are prevalent in various contexts (e.g., achievement assessments and psychological measurements) and offer the advantage of posing multiple questions simultaneously. Zeigenfuse et al. (2020, *Journal of Mathematical Psychology*) introduced an innovative item response model tailored to matching items, addressing unique issues relevant to local dependence and distinct guessing patterns. Despite its pioneering role, this model is not generative, potentially leading to improper item response probabilities. This study advanced this model by revising it into a generative model, thereby facilitating predictive simulations and model evaluations, and enhancing the applicability of the model. In the proposed model, item response functions were derived by marginalizing binary latent variables that represented item-by-item knowledge. However, the marginalization of each person and step can be time-consuming. To reduce the computational load, we proposed an efficient algorithm that was implemented for estimation using R and Stan. A simulation was performed to compare the original and proposed methods as concerns the parameter recovery and calculation time, and the utility of the proposed method was demonstrated.

An Explanatory Hyperbolic Cosine Model with Categorical Covariates

Wednesday, 17th July - 13:45: Topics in IRT 1 (NB A) - Oral

Dr. Jue Wang (University of Science and Technology of China), Prof. George Engelhard (University of Georgia), Prof. Cengiz Zopluoglu (University of Oregon)

Unfolding models, or ideal-point item response models, are commonly utilized for attitude measurement in social sciences. While the inclusion of covariates can enhance the estimation of item response models for cumulative responses, the covariates' effects on ideal-point models that are created for analyzing unfolding responses are not yet fully investigated, particularly with categorical covariates such as grouping variables for individuals. This study proposes an explanatory hyperbolic cosine model that includes categorical person covariates. The model is estimated using a Bayesian method with the Hamiltonian Monte Carlo algorithm. We evaluate the model performance through simulated and empirical data analyses.

The simulation study investigates the accuracy and efficiency of parameter estimation under different conditions. These conditions include varying numbers of covariates (0, 2, 4), types of covariates (binary or non-binary), effect sizes of covariates (0.1, 0.3, 0.5), and different sample sizes (100, 500, 1000). The empirical data analyzed in this study consists of 586 responses to 20 items on attitudes toward censorship. This data also contains information on five categorical covariates, including gender, age group, race, education, and political affiliation. A non-parametric unfolding method is used to select items that fit an unfolding response pattern. We then conduct the analyses based on the selected items using the Stan package in R.

The results of simulation and empirical analyses will be presented at the conference. Implications of explanatory unfolding models with person covariates for attitude measurement will also be discussed.

A Taxonomy of Unfolding Models

Wednesday, 17th July - 14:00: Topics in IRT 1 (NB A) - Oral

Prof. George Engelhard (University of Georgia), Dr. Jue Wang (University of Science and Technology of China)

The purpose of this study is to describe a taxonomy for the classification of unfolding models. Unfolding models (ideal point models) offer an alternative measurement paradigm based on a non-cumulative response process. There is a hodgepodge of models that may be confusing for practitioners, and the proposed taxonomy offers a category system to bring clarity to our understanding of the variety of unfolding models.

The concept of an unfolding response process can be traced back more than 90 years to Thurstone (1928). Thurstone distinguished between increasing-probability (cumulative) and maximum-probability (unfolding) scale types. Current models for unfolding include nonparametric unfolding model (van Schuur, 1984, 1993) and parametric e.g., Generalized graded unfolding model (Roberts, et al., 2000) and hyperbolic cosine unfolding model (Andrich, 1993).

A brief history of unfolding models is presented that includes a discussion of the detailed principles that undergird different models. Next, a Web of Science search is conducted to identify trends and major milestones in research and practice related to unfolding models. Unfolding models are described in detail with key features and principles identified for developing the taxonomy. These principles provide a way to distinguish the unfolding models. The taxonomy also includes a summary of applications of unfolding models

In summary, this study provides a taxonomy and conceptual framework for examining the wide array of unfolding models. The implications for future research and practice are discussed.

Formal theories of psychological phenomena

Wednesday, 17th July - 13:00: Symposium: Formal theories of psychological phenomena (NB B) - Symposium Overview

Mr. Jesse Boot (University of Amsterdam)

Recent years have seen a surge in the development of formal psychological theories. Our symposium features a series of talks that demonstrate how such theories can deepen our understanding of complex psychological processes and discuss how they can be developed using empirical data. Jonas Haslbeck presents a formal theory of both panic disorder and a Cognitive Behavioral Therapy treatment, which demonstrates the potential of formal theories to better understand treatment mechanisms, develop better treatments, and select optimal treatments for (groups of) individuals. Jill de Ron presents a formal model, inspired by ecological models, for cognitive development. The model combines mutualistic relationships and resource competition among cognitive abilities to account for observed developmental phenomena. Also drawing on ecological models, Jesse Boot presents a non-linear dynamical modelling framework for addiction that integrates individual decision-making processes and social dynamics. Bridging the gap between two previously largely disjointed modelling traditions in addiction literature. Finally Denny Borsboom discusses how these formal models shed a different light on the question of how results from intra- and interindividual statistical analyses may be related. Collectively, our talks highlight the versatility and power of formal models in advancing psychological research and theory building.

Improving Treatments for Mental Disorders using Computational Models

Wednesday, 17th July - 13:00: Symposium: Formal theories of psychological phenomena (NB B) - Symposia

Dr. Jonas Haslbeck (Maastricht University), Oisín Ryan (University Medical Center Utrecht), Don Robinaugh (Northeastern University)

Progress in the treatment of psychopathology has slowed and much remains unknown about how treatments achieve their beneficial effects. We propose that computational models can be used to provide new insights into how treatments work and how they can be improved. We argue that treatments are best understood as interventions on systems of interacting components, and that computational models are needed if we are to accurately and precisely determine the effect an intervention will have on this system. We demonstrate this approach by using a computational model of panic disorder to conduct an in silico dismantling study of cognitive behavioral therapy. This simulated trial allows us to: identify a common source of treatment failure; propose a revised treatment protocol that mitigates this source of failure; and demonstrate that, if the model is accurate, this revised protocol will lead to improved treatment outcomes for 10% of patients. We conclude with a discussion of the promise and challenges of using computational models for treatment research.

From Budworms to Behaviors: Bridging the Gap Between Individual Psychology and Social Contexts in Addiction

Wednesday, 17th July - 13:00: Symposium: Formal theories of psychological phenomena (NB B) - Symposia

Mr. Jesse Boot (University of Amsterdam), Mr. Maarten Van den Ende (University of Amsterdam), Prof. Han L. J. van der Maas (University of Amsterdam)

Currently, formal models of addiction either focus on the complex individual decision-making processes involved in addiction or focus on the social dynamics of addiction, treating consumption as a viral entity that spreads across the population. However, current models often fail to integrate these two levels, which has been identified as a key shortcoming of current formal models of addiction. To address this, we propose a non-linear dynamical modeling framework of addiction which integrates both the individual level and social level, striving for a balance between simplicity and empirical relevance, in both. The individual level of our modeling framework is strongly informed by the dual-process theory, which distinguishes between impulsive, automatic actions and controlled, deliberate decision-making. To formalize this theory, we used a well-studied model from ecology, originally used to model periodic outbreaks of the spruce budworm population. At the social level, our modeling framework incorporates the critical processes of selection homophily and peer influence. We show that our integrated modeling framework can be used to explain key phenomena identified in addiction literature on both the individual - and the social level. Moreover, we show how our modeling framework can be extended to include mutualistic, competitive, and more complex interactions between different addictive behaviors.

Towards a general modeling framework of resource competition in cognitive development

Wednesday, 17th July - 13:00: Symposium: Formal theories of psychological phenomena (NB B) - Symposia

Ms. Jill de Ron (University of Amsterdam), Dr. Marie Deserno (University of Amsterdam), Don Robinaugh (Northeastern University), Prof. Denny Borsboom (University of Amsterdam), Prof. Han L. J. van der Maas (University of Amsterdam)

One of the most robust phenomena in cognitive science is the positive manifold, i.e., individuals who score high on one cognitive task tend to score high on other cognitive tasks. The positive manifold has traditionally been explained by the general intelligence (g-) factor. More recently, it has been proposed that the positive manifold arises from mutualistic interactions among cognitive abilities. However, the positive manifold is not uniform across the population: Some cases of atypical development are characterized by specific deficits or uneven cognitive profiles. This poses a challenge to current explanations. In this talk, we propose that competition for limited resources, such as time and environmental factors, is a formative force in cognitive development that can explain these differences in phenotypes. We present a mathematical model combining mutualistic relationships between cognitive abilities with resource competition. The mathematical model is derived from the ecology literature, where resource competition is well documented and modeled, especially between species in a shared environment. We show that the extended model explains positive diversity and phenomena such as developmental stages, slower cognitive development in atypical cohorts, and the absence of early indicators of atypical development. We conclude with an empirical and theoretical research agenda and broader applicability as competition for limited resources is likely to play a role in other areas of psychology, such as psychopathology, in which one symptom consumes resources and causes another symptom to emerge.

Integrating intraindividual and interindividual phenomena in psychological theories

Wednesday, 17th July - 13:00: Symposium: Formal theories of psychological phenomena (NB B) - Symposia

Prof. Denny Borsboom (University of Amsterdam), Jonas Haslbeck (University of Maastricht)

Psychological science is divided into two distinct methodological traditions. One tradition seeks to understand how people function at the individual level, while the other seeks to understand how people differ from each other. Methodologies that have grown out of these traditions typically rely on different sources of data. While both use statistical models to understand the structure of the data, and these models are often similar, research by Peter Molenaar and others showed that results from one type of analysis rarely transfer to the other, unless unrealistic assumptions hold. This raises the question how we may integrate these approaches. In this paper, we argue that formalized theories can be used to connect intra- and interindividual levels of analysis. This connection is indirect, in the sense that the relationship between theory and data is best understood through the intermediate level of phenomena: robust statistical patterns in empirical data. To illustrate this, we introduce a distinction between intra- and interindividual phenomena, and argue that many psychological theories will have implications for both types of phenomena. Formalization provides us with a methodological tool for investigating what kinds of intra- and interindividual phenomena we should expect to find if the theory under consideration were true.

Geometry and psychometrics

Wednesday, 17th July - 13:00: Novel Approaches to IRT (NB C) - Oral

Prof. Francis Tuerlinckx (KU Leuven)

As psychometricians, we use statistical models to measure and possibly explain psychological processes. Geometry and measurement are at least etymologically related, and geometry is deeply rooted in the measurement of quite a number of physical quantities. However, from the author's perspective, geometry has somewhat faded into the background of current psychological measurement and modeling. In this talk, I will not go into the reasons why this may be the case, but I want to give an idea of how ideas from non-Euclidean geometry may be relevant for psychological modeling. We will discuss distance, curvature, and volume in specific applications.

Regularized Gaussian variational estimation for detecting intersectional differential item functioning

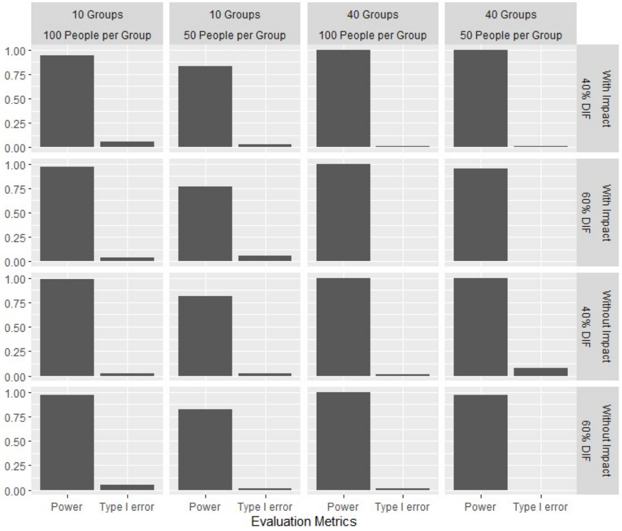
Wednesday, 17th July - 13:15: Novel Approaches to IRT (NB C) - Oral

Mr. He Ren (University of Washington), Dr. Weicong Lvy (University of Washington), Chun Wang (University of Washington), Dr. Gongjun Xu (University of Michigan)

Differential Item Functioning (DIF) occurs when individuals from distinct subgroups differ in the probability of correctly answering an item after controlling their overall performance. While most traditional DIF studies examine subgroups by single demographic, the emerging intersectional DIF emphasizes the compound effect of each person's multiple identities. Existing studies often model DIF by fixed effect and face significant challenges when it comes to intersectionality since they (1) overlook compound effects across demographics and (2) cannot handle numerous subgroups due to cross-combination of demographics.

In this study, we utilize Item Response Theory (IRT) models with random item effect to model DIF, allowing item difficulty to vary across subgroups. Each item difficulty follows a normal distribution with mean as its overall difficulty. A non-zero item-specific variance indicates DIF on that item.

Random item effect models are not new in psychological and educational assessments (De Boeck, 2008; Lathrop & Cheng, 2017; Rijmen & Jeon, 2013). However, prior studies cannot be directly applied to detect intersectional DIF due to difficulties in obtaining exact-zero variance estimates and the computational intensity of existing algorithms. To address these issues, this study (1) introduces log-penalty on item variance, shrinking the variance of DIF-free items to zero and (2) proposes a novel Gaussian variational algorithm that is computationally efficient. Simulations were conducted under different conditions (i.e., number of groups, sample-size per group, DIF item proportion, and impact). As shown in the attached figure, the type I error and power of our method range in [0, 0.08] [0.80, 1], respectively.



Results.jpeg

Bayesian diagnostic classification using doubly bounded visual analogue scaling

Wednesday, 17th July - 13:30: Novel Approaches to IRT (NB C) - Oral

Ms. Hsin Kao (National Taiwan Normal University), Dr. Chen-Wei Liu (National Taiwan Normal University)

Visual analogue scaling (VAS) allows examinees to express their tendencies such as depression, anxiety, and agreement, on a doubly bounded interval scale. However, the use of VAS on diagnostically classifying respondent's latent attributes has not been investigated in the literature. This proposal aims to extend the binary-valued diagnostic classification models (DCMs) to utilize the interval property of the VAS to enhance the person's latent attribute estimation. A novel beta diagnostic classification model (BDCM) is proposed to deal with the doubly bounded VAS item responses, which offers higher estimation accuracy of latent attributes than binary-valued DCMs and enables diagnostic purposes such as recommending career selection for career planning. The proficiency of the BDCM is assessed by a simulation, manipulating number of attributes, sample sizes, and model comparison (i.e., fitting BDCM to VAS data vs. fitting LCDM to dichotomized data from VAS data). The Bayesian Markov chain Monte Carlo algorithms are used to estimate the model parameters. In the empirical study, the BDCM is used to analvze VAS career interest scale based on the Holland Code. The examination of trait classification among examinees and model fit is also incorporated. The results show that BDCM yielded good parameter recovery and classification accuracy, suggesting that it is a prospective statistical methodology for diagnostic classification.

Keywords: visual analogue scale, diagnostic classification model, continuous data, Markov Chain Monte Carlo

Likelihood-free estimation of IRT models in small samples: A neural networks approach

Wednesday, 17th July - 13:45: Novel Approaches to IRT (NB C) - Oral

Dr. Dmitry Belov (Law School Admission Council), Prof. Oliver Lüdtke (IPN – Leibniz Institute for Science and Mathematics Education), Dr. Esther Ulitzsch (Centre for Educational Measurement (CEMO), University of Oslo)

Existing estimators of parameters of item response theory (IRT) models exploit the likelihood function. In small samples, however, the IRT likelihood oftentimes contains little informative value, potentially resulting in biased and/or unstable parameter estimates and large standard errors. To facilitate small-sample IRT estimation, we introduce a novel approach for small-sample IRT estimation that does not rely on the likelihood. Our estimation approach capitalizes on item pool information and trains a neural network (NN) to interpolate the relationship between response patterns and item parameters. We describe and evaluate our approach for the three-parameter logistic (3PL) model; however, it is applicable to any model with an item characteristic curve (ICC). Three types of NNs are developed, supporting to obtain both point estimates and confidence intervals for IRT model parameters. The results of a simulation study demonstrated that for sample sizes of 300 and below, these NNs can perform at the level of Bayesian estimation using Markov chain Monte Carlo (MCMC) methods or even better in terms of quality of the point estimates and confidence intervals, but much faster. Among others, these properties facilitate (1) to pretest items in a real-time testing environment (e.g., CAT), (2) to pretest more items and, as a consequence, to replenish operational pools quicker and assemble more tests as well as (3) to pretest items only in a secured environment (e.g., in strictly proctored test centers) to eradicate possible compromise of new items in online testing.

A differential equation framework for the derivation of item response functions

Wednesday, 17th July - 14:00: Novel Approaches to IRT (NB C) - Oral

Prof. Yvonnick Noel (University of Rennes 2)

Most item response functions at the core of current item response models have been chosen for mathematical convenience. The logistic for example, sometimes justified as a convenient Gaussian CDF proxy, has been extensively used for its nice mathematical properties. In this talk, we present a constructivist approach, where response production is analyzed as the conjugate result of a set of (potentially conflicting) forces, which dynamics is formulated as a differential equation system. Upon integrating the system, an analytical expression for the response function is derived which has a straightforward psychological interpretation. Response functions of the standard cumulative logistic models (1PL, 2PL, 3PL), unfolding models (HCM, Andrich & Luo, 1993; GUM, Roberts & Laughlin, 1996; BUM, Noel, 2014), and asymmetric cumulative models (LPEM, Samejima, 1995), are derived within this framework, using only simple hypotheses of inductive and inhibitive relationships between latent processes. Under this approach, it is straightforward to go beyond first-order (cumulative) and second-order (unfolding) response functions, and derive higher order, potentially asymmetric, functions. Some examples of new models are presented. This approach also offers a new ground for the analysis of item complexity (Bolt & Liao, 2022).

Quantifying replication success: Correspondence measures for replication studies

Wednesday, 17th July - 13:00: Symposium: Quantifying replication success: Correspondence measures for replication studies (NB D) - Symposium Overview

Dr. Vivian Wong (University of Virginia), Prof. Steffi Pohl (Freie Universität Berlin), Dr. Marie-Ann Sengewald (Leibniz Institute for Educational Trajectories, Bamberg), Peter M Steiner (University of Maryland)

The replication crisis in psychology, underscored by the low reproducibility of many study findings, highlights the variability of replication rates based on the chosen correspondence measures. Yet, there is a lack of clear guidelines for selecting the appropriate correspondence measure for determining replication success. In current replication efforts, researchers often do not employ any specific correspondence measure, or they merely assess whether the replicated study's effects are statistically significant. In recent years, new correspondence measures for assessing replication success have been introduced, each with its unique advantages and underlying assumptions. These measures differ in their objectives, assumptions about the significance of effects in the original study, and the potential influence of questionable research practices or publication biases. Most of these measures are designed for post-hoc replications, where a study is replicated after its completion, typically by a different team. There are fewer measures designed for prospective replications, in which the original and replication studies are planned concurrently, usually by the same research team. This symposium proposes new measures for prospective and post-hoc replications, and evaluates existing and new correspondence measures for assessing replication success across different scenarios. The application of the measures is also shown using empirical data. Recommendations will be made regarding the selection of correspondence measures for specific studies.

Implementation of the correspondence test as an S-curve

Wednesday, 17th July - 13:15: Symposium: Quantifying replication success: Correspondence measures for replication studies (NB D) - Symposia

Patrick Sheehan (University of Maryland)

As replication has grown in importance and prominence in the social sciences, the need for methods for assessing replication success has grown. The Correspondence Test (Tryon & Lewis, 2008; Steiner et al. 2023) shows promise over more common methods: it more directly probes the question of replication than simple comparison of the significance pattern of findings, and is a more severe test than either the test of difference or equivalence of effects. However, a key drawback of the Correspondence Test is the current framework is based on two dichotomous null hypothesis significance tests (NHST). Given the concern that overuse of dichotomous NHST has contributed to the replication crisis and the problematic nature of dichotomous hypothesis tests in general (Greenland et al. 2016), it is undesirable that correspondence metrics for replication require the use of NHST. This presentation proposes a generalized implementation of the Correspondence Test as an S-curve (Rafi & Greenland, 2020), a descriptive measure of the totality of the evidence against all possible null hypotheses. S-curves provide a more nuanced view of the evidence provided by the data than is provided by the typical NHST framework. Additionally, the Correspondence S-curve will be compared to the Bayesian Region of Practical Equivalence (ROPE) metric. ROPEs are an alternative metric for assessing correspondence of effect estimates, and provide a similarly nuanced view of replication evidence from a Bayesian, rather than Frequentist, perspective. The two methods are compared under various conditions representing potential replication scenarios.

How best to quantify replication success?

Wednesday, 17th July - 13:30: Symposium: Quantifying replication success: Correspondence measures for replication studies (NB D) - Symposia

Prof. Don van Ravenzwaaij (University of Groningen), Ms. Jasmine Muradchanian (University of Groningen), Dr. Rink Hoekstra (University of Groningen), Prof. Henk Kiers (University of Groningen)

To overcome the frequently debated crisis of confidence, replicating studies is becoming increasingly more common. Multiple frequentist and Bayesian measures have been proposed to evaluate whether a replication is successful, but little is known about which method best captures replication success. In the first part of this talk, I present an attempt to compare a number of quantitative measures of replication success with respect to their ability to draw the correct inference when the underlying truth is known, while taking publication bias into account. Our results show that Bayesian metrics seem to slightly outperform frequentist metrics across the board. Generally, meta-analytic approaches seem to slightly outperform metrics that evaluate single studies, except in the scenario of extreme publication bias, where this pattern reverses. In the second part of this talk, I examine the suitability of meta-analysis specifically to quantify replication success. For a number of original studies, the probability of replication success was calculated using meta-analysis under different assumptions of the underlying population effect when replication results were unknown. The accuracy of the predicted overall replication success was evaluated once replication results became available using adjusted Brier scores. Our results show that meta-analysis performed poorly when used as a replication success metric. In many cases, quantifying replication success using meta-analysis resulted in the conclusion where the replication was deemed a success regardless of the results of the replication study.

A systematic comparison of correspondence measures for prospective replications

Wednesday, 17th July - 13:45: Symposium: Quantifying replication success: Correspondence measures for replication studies (NB D) - Symposia

Dennis Kondzic (Freie Universität Berlin), Jerome Hoffmann (Leibniz Institute for Educational Trajectories, Bamberg), Dr. Marie-Ann Sengewald (Leibniz Institute for Educational Trajectories, Bamberg), Prof. Steffi Pohl (Freie Universität Berlin)

In the last years, the replication crisis has led to a surge of interest in replication studies across various research fields. To assess replication success, different correspondence measures have been developed. Most correspondence measures were developed for post-hoc replications, in which it is assumed that the primary study is already published and sometimes also that the primary study shows a certain result. Correspondence measures have also mainly examined the performance of correspondence measures for post-hoc replications. However, recent research emphasizes the value of prospective replications, in which both studies are planned simultaneously, with the goal of inferring which study characteristics cause effect heterogeneity between both studies. So far correspondence measures have hardly been developed or evaluated for prospective replications. The focus of this study is on providing correspondence measures and guidelines on choosing between them for prospective replications. We present a taxonomy of correspondence measures based on the research questions they answer. We evaluated the existing frequentist and Bayesian correspondence measures for their applicability for prospective replications and adapted some existing measures from post-hoc replications for use in prospective replications. Based on the taxonomy, we evaluated the performance of the measures that aim at the same research question in simulation studies. Our results indicate that some measures make implausibly strong assumptions. Recent measures outperform traditional ones in various conditions. We emphasize the importance of clearly defining research goals for replications. We highlight the assumptions and strength of different measures and provide recommendations for choosing a respective measure.

Assessing replication success over multiple replication study results

Wednesday, 17th July - 14:00: Symposium: Quantifying replication success: Correspondence measures for replication studies (NB D) - Symposia

Dr. Vivian Wong (University of Virginia), Mr. Steffen Erickson (University of Virginia), Dr. Julie Cohen (University of Virginia)

We present results from five experimental studies aimed at assessing the replicability of coaching effects on teacher candidates' pedagogical practices, considering four theoretically informed sources of variation. Our series of replication designs include: a multiple cohort design to evaluate temporal stability, a switching replication design to test robustness across teaching tasks, a multi-modal delivery analysis (online vs. in-person), and a matched conceptual replication design to assess replicability of effects across diverse populations and contexts. For each pairwise comparison of results, we report replication success using multiple correspondence measures that include the magnitude of effects, the sign of effects, the statistical significance patterns, the difference in effects, and the correspondence test results (with tolerance thresholds of 0.2 SD and 1.0 SD). Across most (but not all) measures of correspondence, we find that coaching effects are replicated across variations in timing, task, and delivery modality. However, we find that coaching effects failed to replicate over different setting characteristics, even after adjusting for observed differences in participant characteristics. Results from this presentation will highlight multiple important methodological implications for the analysis of replication efforts – including the need for researchers to pre-specify causal estimands of interest and the correspondence measures that will be used for assessing replication success, as well as the need for adequately powered studies to be included in replication efforts. The presentation concludes by suggesting alternative measures for assessing replication success when more than two studies are included.

Estimating latent variance in variational autoencoders parameterization of Rasch models

Wednesday, 17th July - 13:00: Topics in Statistical/Machine Learning (RB 209) - Oral

Mr. Denis Federiakin (Johannes Gutenberg University of Mainz), Prof. Lidia Dobria (University of Illinois at Chicago), Prof. Olga Zlatkin-Troitschanskaia (Johannes Gutenberg University of Mainz)

Recently, a neural-network-based parameterization of Item Response Theory (IRT) models has been suggested. It utilizes the autoencoder architecture of neural networks and is grounded in the variational inference approach for estimating model parameters. The variational autoencoders approximate the posterior distribution of persons' abilities by a normal distribution, the parameters of which are inferred by the encoding network. Typically, Importance-Weighted Variational AutoEncoders (IW-VAE) are used for this purpose. IW-VAEs take multiple draws from the posterior distribution and weight their reconstruction loss proportionally to its values to smooth the approximation of the target posterior and reconstruct the gradient of the model. However, another approach—FisherNet Variational AutoEncoders (FN-VAE)—can also be used for this purpose. This approach leverages the Fisher Information entity from IRT models to estimate the posterior. Specifically, in this architecture, the variance of the observation-specific standard distribution is calculated as Fisher Information about the draw from the posterior using the decoder (item) parameters. To date, almost all studies in IRT-VAE have focused on analyzing 2-parameter models. The goal of this presentation is to compare the performance of IW-VAE and FN-VAE in estimating parameters of the Rasch models, where item discrimination is fixed to unity and latent variance is estimated. The results of simulations and a real data example of German higher education students' economic literacy assessment show that both FN-VAE and IW-VAE can estimate the latent variance comparably well, with FN-VAE exhibiting a higher overestimation bias.

Unsupervised detection of random responding for Likert-type inventories with varying numbers of response categories

Wednesday, 17th July - 13:15: Topics in Statistical/Machine Learning (RB 209) - Oral

Mr. Michael John Ilagan (McGill University), Dr. Carl Falk (McGill University)

Likert-type inventories administered online risk “random” responding, such as by bots. To safeguard data quality, we consider unsupervised classification of random vs. non-random responders. Previous work proposed a classifier based on a permutation test with bias-corrected outlier statistics. Such a classifier successfully calibrates sensitivity, assuming that for random responders, exchangeability holds for the entire response vector. However, when the items do not have the same number of response categories, the same assumption is invalid. To extend the classifier to inventories with varying number of response categories, we propose grouping items by number of response categories, generating the empirical null distribution by permuting only within each group, otherwise following the same approach. Such a proposal is in contrast to doing a permutation test per item group then combining multiple p-values into a final predicted class. In a simulation study, the main findings were twofold. First, the proposed approach generally outperformed alternatives considered in terms of sensitivity calibration and accuracy. Second, p-values from groups with few items failed to calibrate sensitivity to a 95% nominal rate. An R package, in development, for random responder detection will incorporate this proposed approach.

Predicting Missing Response with BERT Model in Process Data

Wednesday, 17th July - 13:30: Topics in Statistical/Machine Learning (RB 209) - Oral

Dr. Qiwei He (Georgetown University), Mr. Sibo Dong (Georgetown University)

Missing values pose a significant challenge, undermining the integrity and reliability of subsequent studies. Traditional methodologies have predominantly focused on statistical or Machine Learning techniques that rely on feature extraction to mitigate these issues. However, the effectiveness of these approaches is heavily contingent upon the careful selection and engineering of features, thereby limiting the scalability and adaptability of such methods.

Recent advancements in deep learning, particularly in the domains of natural language processing and computer vision, have showcased the transformative potential of Large Language Models (LLMs) in extracting complex patterns and generating insights from unstructured data.

This study proposes Bidirectional Encoder Representations from Transformers (BERT) method to predict missing values in interactive items with process data. Specifically, we trained the BERT model with an input of action sequences and response data and output the correct/incorrect values in the missing response. Different sampling techniques were used to address the sample imbalance issue, especially in items with low percentage of correctness. A total of 11,265 respondents in the PIAAC PSTRE who had both process and response data were used in this pilot study. The BERT with Naive Oversampling strategy shows the best result by increasing approximately 30% in F1 score for high-difficulty items and an overall 4% increase on other items.

Our findings suggest that integrating deep learning, particularly BERT and LLMs, into sequential process data analysis represents a promising direction for future research, with the potential to significantly improve the accuracy in measurement and interactive item design.

Deep-Learning Methods for Multiple Imputation in Large-Scale Survey Assessments

Wednesday, 17th July - 13:45: Topics in Statistical/Machine Learning (RB 209) - Oral

Dr. Usama Ali (Educational Testing Service), Dr. Peter van Rijn (ETS Global), Dr. Paul Jewsbury (Educational Testing Service)

In large-scale survey assessment (LSA), plausible values (PVs) are multiple imputations from students' posterior distributions of proficiency given their item responses and background variables under the item response theory (IRT) model. As an alternative to PVs, market-basket reporting has been explored for assessments like NAEP (Mislevy, 1998) and PISA (Zwitser et al., 2017), where item responses missing by design are imputed using IRT. However, recent advancements in deep-learning methods for imputation (e.g., Lall & Robinson, 2021) offer the potential to reduce reliance on background variables and weaken the dependency on IRT assumptions. We investigate whether deep-learning methods can serve as a robust alternative to current PV methodology in LSA. Comparisons between recent deep-learning imputation and established PV methods used in LSA are crucial. Challenges in machine-learning methods include mimicking IRT methods (e.g., a 2PL model is estimated; Urban & Bauer, 2021), and the strong assumptions of IRT models (e.g., shape of item response functions, local independence, absence of position effects, measurement invariance, unidimensionality). Not accounting for non-responses and non-effortful responses, and including an excessive number of background variables in the model, contribute to the challenges as well. We investigate whether a deep-imputation approach that focuses on prediction of observed data (e.g., through overimputation) can better deal with these issues. Comparisons will be conducted both on a theoretical level and an applied level using LSA data.

Exploring Classic Machine Learning Models and Large Language Models in Detecting ChatGPT Generated Essays in Writing Assessments

Wednesday, 17th July - 14:00: Topics in Statistical/Machine Learning (RB 209) - Oral

Mr. Haowei Hua (The Culver Academies), Mr. Yao Jiayu (Anhui Polytechnic University)

ChatGPT, a powerful generative AI, holds significant role for enhancing K-12 education by offering support in various tasks such as answering questions, solving math problems, and generating content like essays, code, and presentation slides. While it presents an invaluable resource for learning, concerns arise regarding its potential misuse by students for completing school assignments. Current commercial detectors, like Gammarly and GPTZero, are designed for general text generated by AI, lacking specificity for high-stakes assessments. This study addresses the challenge of detecting potential academic cheating using ChatGPT in high-stakes assessments. Classical machine learning methods, including logistic regression, naive Bayes, and decision trees, were employed to identify distinctions between essays generated by ChatGPT and those authored by students. Additionally, pretrained language models such as Roberta and BERT were compared against traditional machine learning approaches. The analysis focused on the prompt 1 from the ASAP Kaggle competition. To evaluate the effectiveness of the detection methods, four approaches were applied to revise ChatGPT-generated essays: Grammarly Premium, revisions by eighth-grade students, revisions by ninth-grade or above students, and further modifications by ChatGPT with additional prompting to humanize and naturalize the essays by introducing grammatical mistakes. In the detection of unmodified ChatGPT essays, Electra, a pretrained language model, demonstrated a high Quadratic Weighted Kappa (QWK) score of 97%, while Support Vector Machine (SVM) outperformed the large language models with a remarkable QWK score of 99%. This research contributes to addressing concerns around academic integrity in high-stakes assessments involving generative AI technologies.

Ordinal Cognitive Diagnosis in Nonparametric Framework

Wednesday, 17th July - 13:00: Topics in Clustering (RB 210) - Oral

Prof. Youn Seon Lim (University of Cincinnati)

Cognitive diagnosis models offer insights into whether test-takers have mastered the specific skills, termed "attributes," within a given knowledge domain. These models define distinct proficiency levels based on attribute mastery, determined by how individuals respond to test items. While attributes are commonly seen as either mastered or not, using ordinal attributes can enhance the precision of assessing attribute mastery. Karelitz (2004) introduced the ordered-category attribute coding framework (OCAC) for polytomous attributes. Other approaches to handle ordinal attributes in cognitive diagnosis have been proposed in the literature. But the large number of parameters often created difficulties in fitting these models. In this study, a nonparametric method for cognitive diagnosis is proposed for use with polytomous attributes, called the nonparametric ordinal attributes diagnostic classification (NOADC) method, that is an adaptation of the OCAC framework. The new NOADC method proposed here can be used with various cognitive diagnosis models. It does not require large sample sizes and exhibits computational efficiency. Extensive simulation studies demonstrate its effectiveness in recovering proficiency classes. The NOADC method is also successfully applied to real-world data.

Sparse convex optimal discriminant clustering

Wednesday, 17th July - 13:15: Topics in Clustering (RB 210) - Oral

Ms. Mayu Hiraishi (Doshisha University), Dr. Kensuke Tanioka (Doshisha University), Prof. Hiroshi Yadohisa (Doshisha University)

With developing information technology, large and complex data can be obtained more easily and we have gaining more opportunities to deal with this type of data. Consequently, it is more important to identify cluster structures within such data. Numerous dimension reduction clustering methods have been proposed to detect information within and between clusters. However, the large number of variables in high-dimensional data may make it difficult to interpret the characteristics of each cluster accurately. To deal with this problem, various methods that perform variable selection and dimension reduction simultaneously have been proposed.

In this study, we focus on Sparse optimal discriminant clustering (SODC) proposed by Wang et al. (2016), which is a sparse constraint-based dimension reduction clustering method. SODC, containing sparse constraints, offers advantage that the tuning parameter can be determined by cross-validation. This method has a good accuracy in capturing the true cluster structure and is easy to interpret. However, SODC is a two-step approach, which is not always possible to yield results that optimize the objective function.

Therefore, we propose a new dimensional reduction clustering method that selects variables and give cluster structure to low dimensional scores simultaneously with one objective function based on SODC.

Generalized Bayesian Method for Diagnostic Classification Models

Wednesday, 17th July - 13:30: Topics in Clustering (RB 210) - Oral

Dr. Kazuhiro Yamaguchi (University of Tsukuba), Mr. Yanlong Liu (University of Michigan), Dr. Gongjun Xu (University of Michigan)

Various parameter estimation methods for the DCMs have been actively developed. Parametric and nonparametric estimation methods are commonly used in DCMs. This loss-function-based parameter estimation method for diagnostic classification models was proposed by Ma et al. (2023, Psychometrika) that can uniformly treat various parameter estimation methods. However, loss-function-based methods exhibit certain limitations. For example, these methods only provide point estimates, which may be problematic because we cannot evaluate how point estimates vary by sampling or variations in the estimation. Furthermore, consider prior knowledge and uncertainty of sampling do not consider in the loss-function-based method. This study extends the loss-function-based parameter estimation method for diagnostic classification models to overcome these problems. We integrate the loss-function-based estimation method with the general Bayesian method. The general Bayesian method is appropriate for the situation that the assumed model may not accurately represent the true data-generating process, or the connection between the model parameters and data may not be described via the assumed model, which is known as a M-open situation. We proposed general posterior expression for the diagnostic classification models and established the consistency of attribute mastery patterns of the proposed method. The proposed general Bayesian method is compared in a simulation study and found to be superior to the previous nonparametric diagnostic classification method—a special case of the loss function-based method. Moreover, the proposed method is applied to real data and compared with previous parametric and nonparametric estimation methods.

Modeling examinee heterogeneity within performance classification using generalized low rank models

Wednesday, 17th July - 13:45: Topics in Clustering (RB 210) - Oral

Dr. Sydne McCluskey (private researcher), Dr. Jay Verkuilen (City University of New York), Dr. Magdalen Beiting-Parrish (Federation of American Scientists), Dr. Howard Everson (City University of New York)

Reckase (2023) emphasizes consistency of performance classifications with external information as an important piece of validity evidence in standard setting and other important psychometric tasks. However, it is difficult to gather compelling external validity evidence because the external data are generally high-dimensional and include considerable construct-irrelevant variation. To address this issue, we adapted Udell et al (2016) Generalized Low Rank Modells (GLRMs), which extend traditional low rank matrix approximation techniques such as PCA, *k*-means, and nonnegative matrix factorization to accommodate a wider variety of data types. These models, for example, can be used as exploratory tools to identify patterns of examinee characteristics that relate to proficiency classifications. GLRMs can also help quantify and describe cluster heterogeneity within and between performance classifications. Anomalous cases or clusters can be described by identifying observations not well-modeled by the GLRM. We will demonstrate this approach using high-stakes accountability test data from a large northeastern US state which includes rich student, school, and test information. Clusters of examinees represented in the reduced-dimensional space will be identified, and relationships between the emergent archetypes (analogous to the PCs in PCA) and performance classifications will be investigated. The goal of the research is to perform exploratory analyses to elucidate group differences within and between performance classification, as group differences are an important piece of the external validity argument supporting substantive interpretations of student proficiency based on performance classification.

Multivariate location-scale models for meta-analysis

Wednesday, 17th July - 13:00: Statistical topics (RB 211) - Oral

Dr. Katrin Jansen (University of Münster), Prof. Steffen Nestler (University of Münster)

Often, primary studies that are pooled in a meta-analysis provide information on multiple outcomes of interest. Multivariate meta-analysis is becoming increasingly popular because it allows to analyze these outcomes simultaneously, which is often more efficient than separate, univariate meta-analyses. Furthermore, it can be used to model more complex data structures, such as those occurring in network meta-analysis. However, multivariate meta-analysis models typically assume that between-study variances and correlations are constant across studies. While it is possible to relax this assumption of constant heterogeneity by using location-scale models in univariate meta-analysis, extensions to the multivariate case have not yet been proposed. In this talk, we close this gap by describing a location-scale model for the multivariate setting where both the between-study variances of the different outcomes and the correlations between them can depend on covariates. We present the results of a simulation study that was conducted to evaluate the performance of multivariate location-scale models estimated with maximum likelihood and Bayesian approaches. Based on our findings, we provide recommendations for the use of multivariate location-scale models for meta-analysis in applied research.

Bayesian modeling of data with ceiling or floor effects

Wednesday, 17th July - 13:15: Statistical topics (RB 211) - Oral

Ms. Ruoxuan Li (University of Notre Dame), Dr. Lijuan Wang (University of Notre Dame)

Ceiling or floor effects, indicating substantial proportions of the maximum or minimum score being observed, are prevalent in psychology research. However, guidance on how to handle ceiling/floor effects in commonly used statistical models of psychology is still limited. The Tobit modeling approach has been found to perform well in handling ceiling/floor effects in independent sample t tests, mediation analysis, and growth curve analyses. This study aims to extend the Tobit modeling approach to handle ceiling/floor effects in four other commonly used models of psychology: dependent sample t tests, moderation analysis, factor analysis, and multilevel models. In this talk, we first introduce how to statistically deal with ceiling/floor effects in those models using a Bayesian Tobit modeling approach. We then evaluate the performance of the Bayesian Tobit modeling approach and compare its performance with that from the naive approach of ignoring ceiling/floor effects via simulations. Simulation results demonstrate that the Bayesian Tobit modeling approach provided satisfactory performance (e.g., accurate estimates and well-controlled Type I error rate) in handling ceiling/floor effects in most studied conditions (e.g., even when homogeneity of variance is violated and the ceiling proportion is as high as 50%). We end the talk by discussing the implications of the results and future research directions.

Marginal and conditional posterior predictive p-values for ordinal models

Wednesday, 17th July - 13:30: Statistical topics (RB 211) - Oral

Ms. Ellen Fitzsimmons (University of Missouri), Prof. Edgar Merkle (University of Missouri)

In Bayesian Structural Equation Models (BSEM), there are many popular metrics for model selection, including the Bayes factor, DIC, and the posterior predictive p-value (ppp-value). Although many metrics are available to guide the model selection process, the relative newness of BSEM leaves room to explore methodology for computing these metrics and their behavior for different model and data types. For example, ppp-values for models using ordinal data (a popular data type in the social sciences) are often computed using the "easiest" likelihood to compute – a marginal likelihood for the latent continuous data underlying the ordinal data (this data is y*). There is not research on whether using this likelihood for computing ppp-values produces the most accurate or efficient results, so we examined the accuracy of and computational runtime for ppp-values computed with different types of likelihoods. In addition to a marginal likelihood for y*, we also tested ppp-values computed with a conditional likelihood for y*, a marginal likelihood for observed ordinal data (y), and a conditional likelihood for y. Conditional likelihoods count latent variables as model parameters, while marginal likelihoods do not. This results in conditional ppp-values being more influenced by the observed data and a tendency to indicate better model fit than the marginal. Additionally, computing conditional and marginal likelihoods for y involves repeatedly evaluating the probability of y occurring, which may be more computationally intensive than using y* for likelihoods. These comparisons show which likelihoods offer the most accurate or efficient ppp-value computations for ordinal data.

Use of external moment information via externally informed models

Wednesday, 17th July - 13:45: Statistical topics (RB 211) - Oral

Mr. Martin Jann (Universität Hamburg), Prof. Martin Spiess (Universität Hamburg)

In this presentation, we will introduce the use of externally informed models for statistical inference, which is a frequentist approach that has recently been developed. This approach enables the incorporation of external moment information, such as means, variances, or more complex moment functions of variables, in a more robust manner by taking into account the uncertainty due to estimation, different study designs, or sampling. Instead of single values, external sets are used to represent external information, which leads to the concept of imprecise probabilities, specifically F-probabilities. On the one hand, inference based on F-probabilities covers convex combinations of probability distributions, which can lead to more distributional robustness compared to models that use a single distribution. On the other hand, external information restricts the range of possible values and can thus reduce variance, counteracting the loss of power when using robust statistics. In practice, this enables more reliable use of externally known mean test scores or moments of auxiliary variables from different sources in standard frequentist inference for new data sets. We will discuss this in the context of the (multiple) linear regression model.

The application of a metaheuristic algorithm when developing a test with a pre-defined value of the cut-off score

Wednesday, 17th July - 14:00: Statistical topics (RB 211) - Oral

Lenka Firtova (Scio)

The study focuses on the development of tests whose objective is to classify examinees into several predefined categories. In practice, these may be for example language tests designed to place candidates into categories defined by the Common European Framework of Reference for Languages (A1, A2, B1, B2 etc.). Let us consider a situation where we have an item bank from which we need to construct a non-adaptive test with a pre-defined value of the cut-off score. In addition, the test needs to meet other requirements (satisfactory discrimination, a sufficient variety of item topics etc.). The process of selection of suitable items from the item bank in order to meet all the criteria mentioned above may be formulated as a nonlinear programming problem (specifically, a problem with a nonlinear objective function) with binary variables, where each variable represents an inclusion or a non-inclusion of an item in the resulting test. Such problems are difficult to solve by common optimization methods, so metaheuristic algorithms may prove a better option here. Specifically, the study explores the use of simulated annealing (Kirkpatrick et al., 1993). The algorithm has been adapted to the nature of the problem, incorporating the principles of Measurement Decision Theory, which is based on the Bayes' theorem. The study has been conducted in R, using real data from an English test.

Computational Aspects of Psychometric Methods and Beyond

Wednesday, 17th July - 14:30: Invited Talk (Vencovského aula) - Invited Talk

Dr. Patrícia Martinková (Czech Academy of Sciences and Charles University)

This talk introduces the research expanding upon the topics of the recently published book “Computational Aspects of Psychometric Methods: With R” (Martinková & Hladká, 2023). Focusing first on inter-rater reliability (IRR), we describe a flexible method for assessing heterogeneity in IRR with variance components models (Martinková et al., 2023) and discuss the relationship between the IRR and false positive rate (Bartoš & Martinková, 2024). Furthermore, we introduce innovative approaches for assessing item functioning and detecting heterogeneity in responses to multi-item measurements, proposing new iterative methods (Hladká et al., 2024a, 2024b) and Bayesian estimation algorithms (Pavlech & Martinková, 2024). We also discuss approaches incorporating more complex data, such as item wording (Štěpánek et al., 2023). Finally, we provide an overview of the software implementation, highlighting the ShinyItemAnalysis R package and interactive application (Martinková & Drabinová, 2018) and its new extendability option via add-on modules (Martinková et al., 2024).

Martinková, P., & Hladká, A. (2023). *Computational Aspects of Psychometric Methods: With R.* Chapman and Hall/CRC. https://doi.org/10.1201/9781003054313

Building bridges between Psychometrics Island and Psychology Mainland

Wednesday, 17th July - 14:30: Invited Talk (RB 101) - Invited Talk

Prof. Eiko Fried (Leiden University)

In the last decades, Psychometrics Island has been bustling with creativity and productivity, resulting in beautiful models such as factor and network analysis. These models have arrived via air mail on Psychology Mainland, often with cryptic instruction material and little information about how models ought to be used. Perhaps it is not surprising that much of the applied literature using these models falls short of sound theory building and testing. For instance, researchers commonly conflate statistical and theoretical models, and there is an over-reliance on fit indices to adjudicate between models. Latent theories are common, when authors use implicit beliefs or causal assumptions to guide inferences. Generally, much of the literature looks as if researchers are baking psychometric cakes following the same recipe over and over again, without generating insight or knowledge. In this talk, I discuss these problems with a focus on factor and network models. I conclude that psychometricians ought to be aware of these problems that threaten the validity of findings, and that you play a crucial role to mitigate and prevent them. We need you, not only on Psychometrics Island, but also on Psychology Mainland.

From Network Psychometrics to Bayesian Graphical Modeling

Wednesday, 17th July - 15:45: Early Career Award Talk (Vencovského aula) - Early Career Award Talk

Dr. Maarten Marsman (University of Amsterdam)

The network model has emerged as a new conceptual model for relating observations to theoretical constructs emerged in the psychometric literature: This new approach spread rapidly throughout the psychological sciences, catalyzed by the rapid development of methodology and accompanying software that allowed psychologists to fit these network models to their own data. The early stages of the psychometric network literature were marked by considerable methodological advances, the establishment of formal links between the new network models and classical latent variable, and the revival of classical psychometric debates (e.g., the idiographic vs. nomothetic debate) and divisions (the parallel literatures on structural equation modeling and item response theory modeling).

But as the field has matured, several methodological and conceptual concerns have emerged. I will focus my talk on the two biggest challenges facing the field: ensuring the robustness of network results and developing longitudinal and cross-sectional models that fit the mostly discrete data we see in practice. I aim to address these challenges by using Bayesian model averaging to account for model uncertainty, which is at the heart of robustness concerns, and by using established links to classical psychometric theory to develop network models that fit psychological data. This research will build on previous developments and emphasize the importance of open source and user-friendly software for broader adoption of these methodological advances. In addition, the new modeling solutions inspired by formal links to classical psychometric models will have implications for ongoing debates and divisions within the evolving field of psychometrics.

Psychometric Methods for Analysis of Wellbeing in the Context of Digital Technologies

Thursday, 18th July - 09:00: Symposium: Psychometric Methods for Analysis of Wellbeing in the Context of Digital Technologies (Vencovského aula) - Symposium Overview

Dr. Patrícia Martinková (Institute of Computer Science of the Czech Academy of Sciences)

This symposium explores novel psychometric methods in the evolving landscape of utilizing digital technologies to enhance physical, psychological, and social wellbeing. The symposium opens with an introduction to the DigiWELL project (2024-2028) related to these topics. The first presentation by Misha Pavel will delve into the main challenges in utilizing intensive longitudinal data (ILD) to improve wellbeing, and will offer novel approaches that blend machine learning with mechanistic modeling to capture the dynamics of individuals. The second talk by Oriol J. Bosch addresses the reliability of digital trace data in media exposure measures, employing a multiverse of measurements analysis to enhance understanding of web tracking measures, their quality, and the impact of design choices. The third talk by Young Won Cho focuses on handling non-ignorable missingness in ILD through joint modeling, offering guidance on fitting dynamic models in scenarios with data missing not at random. The fourth presentation by David Lacko explores the utilization of dynamic structural equation modeling on ILD of digital media use, presenting opportunities for communication scholars to examine short-term, person-specific, and reciprocal media effects. The last talk by Jaroslav Hlinka outlines an approach for quantification, testing, treatment, and interpretation of non-gaussian dependences in intensive longitudinal data, discussing lessons learned from complex dynamical systems, particularly the assessment of the role of specific temporal patterns and signal nonstationarities. Throughout the symposium, Steriani Elavsky will serve as the discussant.

Dynamic Systems Modeling of Human Behaviors to Improve Proactive Healthcare

Thursday, 18th July - 09:00: Symposium: Psychometric Methods for Analysis of Wellbeing in the Context of Digital Technologies (Vencovského aula) - Symposia

Prof. Misha Pavel (Northeastern University), Prof. Holly Jimison (Northeastern University)

In this presentation, we consider psychometric approaches to modeling that govern many recent advances in healthcare. Future healthcare is shifting towards predictive, proactive, personalized assessment and therapy. Research findings show that individuals' health-related behaviors and psychological states are critical to their health outcomes and quality of life. Unlike static measurements, e.g., conventional cognitive tests, assessments must incorporate the dynamics of mental states (anxiety, stress) and environmental and social contexts. Recent sensor, computation, and communication technology advances provide an unprecedented opportunity for intensive, longitudinal measurement. This presentation will cover the motivation for mathematical (computational) models combined with longitudinal observations and assessment data. We will outline the major issues using intensive longitudinal data (ILD) to improve participants' activity levels. The central element of our approach involves the development of novel computational modeling and prediction. Temporally dense, longitudinal data are amenable to sequential statistical and engineering signal analysis and systems analysis to make model-based predictions that enable optimal interventions. We will discuss several theoretical frameworks and our research in mechanistic modeling, including systems identification approaches. Starting with the definition of models and their role in measurement, including sampling frequency, we describe several approaches to modeling systems dynamics. We compare popular machine-learning approaches to mechanistic modeling. We will illustrate our approaches using the results of a recent large NIH-funded study of 'participants' activity, such as walking. The participants' data included Fitbit data and periodic Ecological Momentary Assessments.

The Reliability of Digital Trace Data in Media Exposure Measures: A Multiverse of Measurements Analysis

Thursday, 18th July - 09:00: Symposium: Psychometric Methods for Analysis of Wellbeing in the Context of Digital Technologies (Vencovského aula) - Symposia

Dr. Oriol Bosch Jover (University of Oxford)

Understanding online media exposure is critical, both for the study of political phenomena and to understand their effect on people's mental wellbeing. Given the doubts about survey self-reports research on media exposure has turned to web tracking data, sometimes considered the gold standard. However, studies revealed that web tracking data is also biased.

To improve the understanding of the quality of web tracking measures of media exposure, this paper estimates their true-score reliability. It does so by leveraging the longitudinal nature of our dataset, to use the Quasi-Markov Simplex Model. The paper also introduces a new approach to identify the design choices that might optimize the reliability of web tracking measures: the multiverse of measurements analysis. Specifically, using data from a three-wave survey in Spain, Portugal, and Italy, combined with web tracking, this paper conducts a assesses the reliability of +2,500 web tracking measures of media exposure. Results show an overall high, but imperfect, reliability (0.86). Additionally, results suggest that the design decisions made by researchers can have a substantial impact on the quality of the web tracking data. All in all, our results can help researchers better understand the quality of web tracking measures, the mechanisms in which different design decisions can affect it, and more generally approaches to combine psychometrics and computational methods to study the quality of digital trace data more generally.

Joint Modeling for Non-Ignorable Missingness in Intensive Longitudinal Data

Thursday, 18th July - 09:00: Symposium: Psychometric Methods for Analysis of Wellbeing in the Context of Digital Technologies (Vencovského aula) - Symposia

YOUNG WON CHO (The Pennsylvania State University), Dr. Sy-Miin Chow (The Pennsylvania State University)

We aim to provide guidance for handling intermediate missingness in intensive longitudinal data (ILD) by investigating the usability of joint modeling, particularly in scenarios where data are missing not at random. Joint modeling is a technique that simultaneously fits both substantive models and missing mechanism models. While joint modeling has shown promise in handling missingness in longitudinal data in medical applications, its application in ILD collected from wearable devices and with multilevel dynamic models remains underexplored.

We first present an empirical illustration of our proposed approaches using real-world ILD, comprising one year of daily observations on affect and physical activity tracked by wearable devices. A multilevel Vector Autoregressive (VAR) model was used as our substantive model to understand the reciprocal influences between daily affect and physical activity. By comparing the results of joint modeling with those of other common missing data handling approaches, we emphasize the importance of sensitivity analysis for evaluating the robustness of the model estimates and the careful selection of missing mechanism models based on assumed or diagnosed mechanisms.

Subsequently, we assess the effectiveness of joint modeling through simulated data representing real-world ILD conditions, such as person-specific compliance rates and clustered located missingness patterns. By mirroring realistic ILD missingness scenarios, we evaluate the accuracy of joint modeling in estimating dynamic models and identifying missing mechanisms. Our findings underscore the potential of joint modeling as a valuable tool for addressing non-ignorable missingness in ILD.

Utilizing Dynamic Structural Equation Modeling on Intensive Longitudinal Data: The Inspiration for Digital Media Use Effects Research

Thursday, 18th July - 09:00: Symposium: Psychometric Methods for Analysis of Wellbeing in the Context of Digital Technologies (Vencovského aula) - Symposia

Mr. David Lacko (Institute of Psychology, Czech Academy of Sciences), Jana Blahošová (Interdisciplinary Research Team on Internet and Society, Masaryk University), Michaela Lebediková (Interdisciplinary Research Team on Internet and Society, Masaryk University), Michal Tkaczyk (Interdisciplinary Research Team on Internet and Society, Masaryk University), David Šmahel (Interdisciplinary Research Team on Internet and Society, Masaryk University), Steriani Elavsky (Faculty of Education, University of Ostrava), Martin Tancoš (Interdisciplinary Research Team on Internet and Society, Masaryk University)

Intensive longitudinal data (ILD) on digital media use provides new analytical opportunities for communication scholars concerned with short-term, person-specific, and reciprocal media effects. Recently developed dynamic structural equation modeling (DSEM) offers a unique opportunity to examine this type of data by adopting two-level modeling with time on Level 1 and individuals on Level 2, distinguishing within- and between-person effects via latent person-mean centering. This facilitates the modeling of intraindividual changes over time. However, the efficient utilization of this analytical approach is limited by several conceptual and data-driven constraints. In this study, we demonstrate the possibilities of DSEM in media effect research on ILD obtained from ecological momentary assessment, consisting of four 2-week bursts with daily surveys, along with objective measurements of smartphone use for social and entertainment apps. A total of 17,580 observations were collected from 197 Czech adolescents (ages 13-17). Special emphasis is placed on the practical methodological and statistical challenges inherent in utilizing DSEM for media effect research, such as optimal temporal sequencing of objective and self-reported data, or identification of meaningful and theoretically sound temporal lags for analysis of specific short-term effects (e.g., 15 minutes, 30 minutes, or 60 minutes before and after measurement of affect). Furthermore, possibilities of DSEM for model building and hypotheses crafting in digital media use research are discussed.

Quantification, testing, treatment and interpretation of non-gaussian dependences in intensive longitudinal data: lessons learned from complex dynamical systems

Thursday, 18th July - 09:00: Symposium: Psychometric Methods for Analysis of Wellbeing in the Context of Digital Technologies (Vencovského aula) - Symposia

Dr. Jaroslav Hlinka (Institute of Computer Science of the Czech Academy of Sciences)

Quantification of relations between measured variables of interest by statistical measures of dependence is a common step in analysis of intensive longitudinal data. The choice of dependence measure is key for the results of the subsequent analysis and interpretation. The use of linear Pearson's correlation coefficient is widespread and convenient. On the other side, as behavioral dynamics is widely acknowledged to be a nonlinear system, nonlinear dependence quantification methods, such as those based on information-theoretical concepts, could be argued to be more suitable for this purpose. In this contribution we outline an approach that enables well-informed choice of dependence measure for a given type of data, improving the subsequent interpretation of the results. The presented multi-step approach includes statistical testing, quantification of the specific non-linear contribution to the interaction information by means of extra-normal information (Hlinka et al., 2011, NeuroImage), identifying the variables with the strongest nonlinear contribution to dependences, and assessment of the role of specific temporal patterns, including signal nonstationarities. We demonstrate the method by illustrative examples from applications to other complex systems including brain activity, long-term climate measurements and stock prices, that highlight the possible common or more exotic scenarios of apparent nonlinearities in dependences structure of longitudinal data, particularly stressing the nongaussianity as a marker of various process nonstationarities, and the possible data preprocessing treatments. Finally, the potential applications to intensive longitudinal data in the context of tracking wellbeing are discussed, with particular emphasis on the analysis of actigraphic records and mobile-application-based self-reports.

Discovery of attribute hierarchies in longitudinal cognitive diagnostic models

Thursday, 18th July - 09:00: Longitudinal Latent Class and CDM Models (RB 101) - Oral

Dr. Yinghan Chen (University of Nevada Reno), Dr. Shiyu Wang (University of Georgia)

Attribute hierarchy, the underlying prerequisite relationship among attributes, plays an important role in applying Cognitive Diagnosis Models (CDMs) for designing learning interventions. There is still a lack of efficient statistical tools to directly learn attribute hierarchy from the observed data, especially in a dynamic learning process. We construct a Bayesian framework for dynamic CDMs with hierarchical latent attributes, and propose an efficient Metropolis-within-Gibbs algorithm to estimate the underlying hierarchical structure of skills. In particular, we impose hierarchical constraints in both initial permissible patterns and permissible transitions of attribute profiles. The algorithm is able to discover the underlying structure or substructures and provide a better classification of students' learning trajectories.

Regularized variational Bayes for Q-matrix inference in cognitive diagnosis models

Thursday, 18th July - 09:15: Longitudinal Latent Class and CDM Models (RB 101) - Oral

Ms. YI JIN (The University of Hong Kong), Prof. JINSONG CHEN (The University of Hong Kong)

In the realm of cognitive diagnosis models, accurately mapping the relationship between attributes and items, as calibrated in the Q-matrix, is crucial. Traditional Bayesian statistical methods for estimating the Q-matrix have been widely adopted. However, their dependency on Markov chain Monte Carlo (MCMC) process requires a sufficiently large number of iterations and burn-in periods. Such computational burdens pose great challenges, especially when dealing with complex datasets (e.g., large samples, many attributes, many items). To tackle the issue, this study introduces a compelling alternative: a regularized variational Bayes approximation approach for Q-matrix inference, aiming to enhance computational efficiency without compromising precision. By integrating regularization into the variational Bayes framework, the proposed method can efficiently select the significant model parameters while shrinking those insignificant ones towards zero, which not only promotes sparsity but also facilitates a scalable and efficient estimation.

The main thrust of this study includes: 1) the development of a novel mixture of formulations in saturated forms; 2) the extension of variational Bayesian expectation-maximum (VBEM; Bishop, 2006) through adding regularization penalty terms to simultaneously estimating model parameters and recovering the Q-matrix; 3) a comprehensive comparative analysis between the regularized-variational-Bayes-based method, and MCMC-based method for Q-matrix inference using both simulated and real data.

A Wrinkle in the IRT Theta Continuum: Exploring DCM Classification

Thursday, 18th July - 09:30: Longitudinal Latent Class and CDM Models (RB 101) - Oral

Jonathan Templin (University of Iowa), Sergio Haab (University of Iowa), Jacinta Olson (University of Iowa), Ariel Aloe (University of Iowa)

The person parameters emanating from psychometric models are the basis for test scores given to the assessment respondents. Many, if not most, assessments seek to provide a score along a continuum, often using psychometric models known as item response theory (IRT) models. In practice, however, respondents tend to be classified into categories (e.g., proficient, or not) established by their test score. It is these categorizations, derived from the assessment scores, that are useful for decision making (e.g., which students receive educational interventions, or which respondents are eligible for licensure). Meanwhile, some psychometric models (e.g., diagnostic classification models; DCMs) provide classifications without placing respondents' scores onto a continuum, focusing solely on classifying respondents as being of higher or lower ability compared to the rest of the respondents. Until now, the two modeling families (IRT and DCMs) have been separate. We present a new method that combines IRT and DCMs–applying a novel algorithm to IRT model estimates to yield DCM parameter estimates and respondent classifications. As part of this work, we show that IRT item and person parameters can be used to provide DCM-model parameters and classification via one additional estimation step. Moreover, we establish these results hold regardless of the number of dimensions in a model, provided a comparable item model (i.e., Q-matrix and latent variable interactions) is maintained. Via a simulation study and an empirical data example, we show that DCM classifications are norm-referenced and the classification results are largely determined by the distribution of the IRT latent variable.

Stepwise estimation approaches of complex latent class models: best practices

Thursday, 18th July - 09:45: Longitudinal Latent Class and CDM Models (RB 101) - Oral

Dr. Zsuzsa Bakk (Leiden University)

Latent class (LC) models are used as a tool to create a clustering on a set of observed indicators, when the number of clusters is unknown. These models belong to the family of latent variable models, with both observed and latent variables treated as categorical. The last 20 years are marked by an increased interest in bias-adjusted approaches to stepwise LC modeling. These approaches separate the estimation of the measurement model, defined by the indicators of the LC model, and that of the structural model, that relates the LC's to external variables of interest. In this presentation I introduce the simultaneous, and two bias-adjusted stepwise estimators, namely the two and three-step estimators. I summarize findings from my research to give general recommendations on which estimators to use under different scenario's, namely in simple LC models where all model assumptions hold, models where the underlying model assumptions are violated, and complex LC models applied to multilevel and longitudinal data. I present results of a set of simulation studies that proposes residual statistics for identifying model misfit, and looks into the severity of ignoring misfit with the three different estimators including for large, complex models. I discuss parameter bias, efficiency and robustness of the proposed estimators under the different scenario's. My results show that across the different conditions the bias-adjusted two step estimator is the most robust against misspecifications, as long as the measurement model is strong, while the simultaneous estimator is the most efficient when all model assumptions hold.

Novel statistical approaches to tackle challenges in intensive longitudinal data

Thursday, 18th July - 09:00: Symposium: Novel statistical approaches to tackle challenges in intensive longitudinal data (NB A) - Symposium Overview

Marieke Schreuder (KU Leuven)

Over the past decades, psychological research became increasingly invested in using intensive longitudinal (IL) methods, in which participants provide data (e.g., self-reports) multiple times per day, for several weeks or months. These methods are often motivated by the notion that psychological processes are essentially ideographic and dynamic – meaning that they are unique to individuals and evolve over time. It is precisely this notion that is also challenging our field. Firstly, the focus on ideographic experiences challenges generalization towards a broader population. To overcome this, we need methods that harmonize idiosyncrasies and population-level inferences. One such method will be covered in this symposium, namely advanced subgroup detection. Specifically, **Jody Zhou** will show how regularization can improve the estimation of Gaussian Mixture Models. A second challenge pertaining to modeling IL data concerns the fact that the time-varying nature of psychological processes is often inadequately addressed. Again, this can only be overcome by reconsidering common time series analyses, including autoregressive models and moving window analyses. Along these lines, **Marieke Schreuder** will show how survival analysis may be used to estimate key emotion dynamics. **Evelien Schat** will discuss a person-specific updating approach to deriving exponentially weighted moving average control limits, when too little baseline data are available of the psychological process. Finally, **Fridtjof Petersen** compares methods for choosing optimal training window sizes when predicting self-reported emotions from sensor data. Taken together, this symposium includes four innovative ways of dealing with the conceptual and statistical challenges that characterize IL data in psychology.

Improving the Estimation of Sample Heterogeneity in Multivariate Dynamic Processes through Feature Selection

Thursday, 18th July - 09:00: Symposium: Novel statistical approaches to tackle challenges in intensive longitudinal data (NB A) - Symposia

Ms. Di Jody Zhou (University of California, Davis), Dr. Sebastian Castro-Alvarez (University of California, Davis), Dr. Siwei Liu (University of California, Davis)

Dynamic psychological processes are typically heterogeneous across individuals. Gaussian mixture models (GMM) have become increasingly popular for detecting such heterogeneity. However, applying GMM in the context of dynamic modeling can be challenging because dynamic models are often over-parameterized. As a result, the performance of GMM is likely to deteriorate notably due to many unnecessary parameters that complicate the subgrouping algorithm. Reducing the number of non-informative features has been shown to improve subgrouping accuracy. In this study, we propose integrating regularization, a feature selection technique, into GMM to discard non-informative features in vector autoregressive (VAR) based dynamic modeling. We first introduce the rationale behind this penalized dynamic Gaussian mixture method. We then present simulation results to evaluate the performance of this method compared to a traditional mixture subgrouping method without feature selection under data conditions commonly encountered in studies of dynamic psychological systems.

Investigating emotional recovery using survival analyses of burst ESM data

Thursday, 18th July - 09:00: Symposium: Novel statistical approaches to tackle challenges in intensive longitudinal data (NB A) - Symposia

Marieke Schreuder (KU Leuven), Dr. Sigert Ariens (KU Leuven), Prof. Ginette Lafit (KU Leuven), Prof. Eva Ceulemans (KU Leuven)

Many experience sampling (ESM) studies suggested that high psychological resilience is reflected by quickly recovering one's emotional baseline. However, former studies relied on indirect proxies for emotional recovery, namely autocorrelations. These have the downsides of (1) assuming exponential decay towards the emotional baseline, (2) being highly dependent on sampling frequency, and (3) not differentiating between up- versus downward departures from the baseline (i.e., positive versus negative emotional episodes). This preregistered proof-of-concept study aimed to overcome these limitations applying multilevel survival analyses to high-resolution ESM data. Adults (N=69) participated in a three-week ESM study with eight assessments per day, complemented by short-spaced burst assessments that were triggered upon reporting high or low levels of emotions. Resilience was assessed at baseline (trait-level; TR) and daily (day-level; DR). Multilevel survival analyses showed that high DR predicted faster returns from negative episodes, but also delayed returns following positive episodes (exp(β)=1.36, p=0.003). Instead, high TR predicted faster returns from positive, but not negative, emotional episodes (exp(β)=0.98, p=0.035). These findings remained when accounting for emotion intensity and stress pile-up. This illustrates how innovative ESM designs combined with survival analyses may further our insight into emotion dynamics, such as emotional recovery.

Person-specific updating approach to deriving EWMA control limits in case of few in-control observations

Thursday, 18th July - 09:00: Symposium: Novel statistical approaches to tackle challenges in intensive longitudinal data (NB A) - Symposia

Dr. Evelien Schat (KU Leuven), Prof. Francis Tuerlinckx (KU Leuven), Prof. Eva Ceulemans (KU Leuven)

Retrospective analyses of experience sampling (ESM) data have shown that changes in mean levels may serve as early warning signals of an imminent depression. Detecting such early warning signs prospectively would pave the way for timely intervention and prevention. The exponentially weighted moving average (EWMA) procedure seems a promising method to scan ESM data for the presence of mean changes in real time. First, this procedure captures the natural variation present in a set of in-control data, used to establish control limits. Afterwards, incoming data are compared to the in-control distribution, to detect and test whether and when the incoming data go out-of-control (i.e., when the data go beyond the control limits). One of the biggest challenges of applying the EWMA procedure to ESM data, is the amount of in-control data that is needed for optimal performance, which amounts to at least 50 days. Clearly, it is not trivial to obtain such a large amount of in-control data of a single person. We therefore investigate whether we can use the person's incoming data to update the control limits over time, thereby circumventing the need for the large initial in-control dataset. This updating approach can result in more accurate control limits, and thus enhance EWMA performance. Based on simulations, we provide insight into the benefits and challenges of this updating approach.

Training Window Selection Methods for Passive Sensing

Thursday, 18th July - 09:00: Symposium: Novel statistical approaches to tackle challenges in intensive longitudinal data (NB A) - Symposia

Fridtjof Petersen (University of Groningen), Laura F. Bringmann (University of Groningen)

The widespread adoption of smartphones creates the possibility to passively monitor everyday behavior via sensors. Sensor data has been linked to moment-to-moment psychological symptoms and mood of individuals and thus could alleviate the burden associated with repeated measurement of symptoms. Additionally, psychological care could be improved by predicting moments of high psychopathology and providing immediate interventions. Current research assumes that the relationship between sensor data and psychological symptoms is constant over time - or changes with a fixed rate: Models are trained on all past data or on a fixed window, without comparing different window sizes with each other. This is problematic as choosing the wrong training window can negatively impact prediction accuracy, especially if the underlying rate of change is varying. As a potential solution we compare different methodologies for choosing the correct window size ranging from treating the window as a hyperparameter to super learning algorithms. In a simulation study we vary both the underlying relationship type (linear, non linear) as well as the relationship form over time (constant, fixed change, varying change). We show that these methods are able to choose the training windows that approximately reduce the prediction error for both simulated and real world data.

Notes toward a grammar of validity arguments

Thursday, 18th July - 09:00: Theoretical and Applied Issues in Validity (NB B) - Oral

Keith Markus (John Jay College of Criminal Justice, CUNY), Prof. Denny Borsboom (University of Amsterdam)

One criticism of Messick's (1989) validity chapter was that it did not provide sufficiently concrete guidance to test developers, something Kane (1992, 2006, 2013) addressed by further expanding on Cronbach's (1988) notion of a validity argument and that Chapelle (2021) has developed further. A grammar specifies how smaller elements of discourse combine to form larger units. Current theory analyzes validity arguments into warrants, assumptions, backing and rebuttals and Chapelle applies these to different inferences. This takes us in the direction of a grammar of validity arguments, something potentially useful to validity theorists and test developers. Such a grammar should cover anything that needs justification. Validity arguments tend to adopt a multiple-hurdle approach to combining different warrants and assumptions and this appears attributable to a narrow focus on one test design option. However, these choices require justification. If we consider different design options within a validity argument, then there are multiple paths to validity. Another under-developed area is the manner in which different forms of backing combine together. Different types of elements can differ in their relations to other elements. Different elements can differ in the grammar of how they combine with others. Different warrants can differ in how assumptions, backing and rebuttals combine to support them. Finally, typical validity arguments assume certain values and evaluate the interpretation or use in relation to them but little guidance is offered for justifying the values themselves. This leads to a potential extension of Messick's proposed contrastive analysis.

Integrating IRT and GT via a four-building-block approach

Thursday, 18th July - 09:15: Theoretical and Applied Issues in Validity (NB B) - Oral

Mr. Mingfeng Xue (University of California, Berkeley), Prof. Mark Wilson (University of California, Berkeley)

This research proposes a novel approach to integrate Item Response Theory (IRT) and Generalizability Theory (GT) through a four-building-block framework (Wilson, 2023) to enhance measure development and illustrate their usefulness through an empirical study.

Construct map

A construct map defines the construct to be measured and outlines several qualitatively different levels/waypoints for the construct, enhancing the interpretation power of IRT and GT.

Item design

Items are the realization of the construct map and are central to IRT and GT. Our approach embodies the idea that items are sampled from different distributions based on the levels/waypoints they assess.

Outcome space

This block categorizes students' responses into the waypoints designed at the construct map, which yielded the observed scores for IRT and GT.

Statistical model

We propose a multilevel IRT model combining IRT and GT. It builds on earlier work by De Boeck (2008) but has a new element: Item difficulties are conceptualized as the sum of a fixed effect representing the level/waypoint that the item assesses plus a random deviation from the fixed effects.

Empirical study

An empirical study about argumentation measurement in science education illustrates our integration. A construct map with four levels guides the item design, resulting in 24 items administered to 930 students. We apply our statistical model to the data and compare the results with conventional IRT models and GT analysis in terms of multiple aspects.

This research advances measurement theory by synergizing IRT and GT, offering a comprehensive framework for measure development.

Mitigating social desirability using Item quadruplets and MIMIC

Thursday, 18th July - 09:30: Theoretical and Applied Issues in Validity (NB B) - Oral

Dr. Felipe Valentini (Graduate School of Psychological Assessment, University São Francisco), Mr. Rafael Valdece Sousa Bastos (Graduate School of Psychological Assessment, University São Francisco)

Self-report is a pivotal method for data collection in psychology and social sciences due to its convenience. However, it's vulnerable to response biases like social desirability, which complicates research estimates, potentially altering variable relationships and instrument dimensionality. Addressing this, the literature has explored psychometric models to mitigate social desirability effects. Peabody's approach involves item quadruplets to adjust for social desirability, though this can yield unnatural item phrasing. We propose a two-step solution: initially, assess and select effective quadruplets to estimate social desirability, then use these estimates to adjust remaining items. Our model's efficacy was tested across various conditions with simulated data, varying: number of quadruplet (1, 3, 6, 12), factor dimensions (two, six), response bias levels (factor loadings of 0.10, 0.25, 0.50), content factor loading fluctuations (0%, 10%, 20%, 40%), baseline content factor loadings (0.4, 0.55, 0.7), and sample sizes (2000, 4000). Comparing true parameters with estimated biases and coverage, results indicated optimal performance with 12 quadruplets in two-factor structure. However, increasing factor numbers showed that even minimal quadruplets could accurately estimate social desirability for additional items. This suggests that despite item construction challenges, external item social desirability can be effectively estimated, enhancing research accuracy.

Assessing Teaching Skills: Overcoming Psychometric Hurdles in Observational Measures

Thursday, 18th July - 09:45: Theoretical and Applied Issues in Validity (NB B) - Oral

Mr. Steffen Erickson (University of Virginia)

This paper examines psychometric phenomena endemic to observational measures used to assess the quality of specific teaching skills. Item scores representing distinct skills tend to correlate highly on observation rubrics (Bartanen et al., 2023), leading to factor analytic models that converge to single dimensions (Kane & Staiger, 2012; VanderWeele & Vansteelandt, 2022). This can result in incorrect conclusions about the dimensionality of observation rubrics, impeding inferences about the development of specific teaching skills and their relation to other measures of teaching effectiveness. The paper addresses three psychometric drivers of this methodological challenge: correlated error structures, causal relations among skills, and mutual dependencies on covariates. To illustrate these issues, the paper uses an applied example of pre-service teachers learning to metacognitively model solving mathematical word problems. Teacher candidates were taught to unpack the context, quantities, and relationships in word problems, and then to articulate their thought processes in understanding the problem. The proficiency of teacher candidates in these skill areas is measured through observations of standardized performance tasks. It is shown that failing to account for additional covariance between item scores from task and rater influences, the dependence of self-instruction on word problem unpacking, and the mutual dependence on mathematical content knowledge leads to incorrect conclusions about unidimensionality in factor analysis. The results highlight the consequences of misspecifying measurement models, showing how incorrect conclusions about relationships between metacognitive modeling and future classroom performance can occur. The presentation emphasizes the role of theory and knowledge of observation procedures in specifying measurement models.

On the Improvement of Predictive Modeling Using Bayesian Stacking and Posterior Predictive Checking-An Example of Modeling Gender Inequality in Reading Using PISA 2018

Thursday, 18th July - 09:00: Bayesian Stacking, Trees, and Predictive Modeling (NB C) - Oral

Dr. Mariana Nold (Friedrich-Schiller-Universität Jena, Faculty of Social and Behavioural Sciences, Institute of Sociology, Carl-Zeiß-Str. 3 07743 Jena , Germany), Dr. Florian Meinfelder (Faculty of Social Sciences, Economics, and Business Administration, Chair of Statistics and Econometrics), Prof. David Kaplan (University of Wisconsin - Madison)

Model uncertainty is pervasive in real world analysis situations and is an often-neglected issue in applied statistics. Standard approaches to the research process do not address the inherent uncertainty in model building and, thus, can lead to overconfident and misleading analysis interpretations. One strategy to incorporate more flexible models is to base inferences on predictive modeling. This approach provides an alternative to existing explanatory models, as inference is focused on the posterior predictive distribution of the response variable. Predictive modeling can advance explanatory ambitions in the social sciences and in addition enrich the understanding of social phenomena under investigation. Bayesian stacking is a methodological approach rooted in Bayesian predictive modeling. In this paper, we outline the method of Bayesian stacking but add to it the approach of posterior predictive checking (PPC) as a means of assessing the predictive quality of those elements of the stacking ensemble that are important to the research question. Thus, we introduce a viable workflow for incorporating PPC into predictive modeling using Bayesian stacking without presuming the existence of a true model. We apply these tools to the PISA 2018 data to investigate potential inequalities in reading competency with respect to gender and socio-economic background. Our empirical example serves as rough guideline for practitioners who want to implement the concepts of predictive modeling and model uncertainty in their work to similar research questions.

Incorporating sparsity into Bayesian stacking procedures

Thursday, 18th July - 09:15: Bayesian Stacking, Trees, and Predictive Modeling (NB C) - Oral

Ms. Kjorte Harra (University of Wisconsin - Madison), Prof. David Kaplan (University of Wisconsin - Madison)

Bayesian stacking is a procedure adopted from machine learning that allows researchers to combine multiple unique models and optimize overall predictions. Bayesian stacking has the added benefit of not relying on strong assumptions necessary for Bayesian model averaging (BMA) (Yao et al., 2018). For individual models, Bayesian regularization methods have demonstrated stronger predictive accuracy than unregularized modeling approaches (Harra & Kaplan, 2023). Inducing sparsity in statistical models decreases variance and optimizes prediction. While model stacking is not intended to serve as a method for performing variable selection, we are unaware of any systematic investigation examining how sparsity-inducing priors applied to member models in a stack could conceivably lead to more accurate predictions. The present work investigates whether the addition of regularization via sparsity-inducing priors of individual member models can be a worthwhile practice when using Bayesian stacking procedures. Our preliminary findings based on real data and a simulation study suggest that inducing sparsity in member models for stacking does not appear to improve predictive performance.

Bayesian posterior predictive checks to analyze flexible distributional models

Thursday, 18th July - 09:30: Bayesian Stacking, Trees, and Predictive Modeling (NB C) - Oral

Ms. Paulina Grekov (University of Wisconsin - Madison), Dr. James Pustejovsky (University of Wisconsin - Madison)

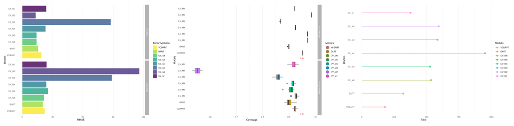
Education research often deals with hierarchical data structures, with research questions pertaining to students or teachers within schools or districts, or to children within clinics. Generalized additive models for location, scale, and shape (GAMLSS) are a flexible class of distributional models that extend on the more restrictive hierarchical linear models (HLM) and generalized linear mixed models. GAMLSS can handle a wide range of response distributions and are unique in their ability to model not only the mean parameter but also the dispersion parameter of the outcome distribution using both fixed and random effects. When using complex, hierarchical models such as GAMLSS, researchers face the challenge of interpreting and communicating findings to a broader audience. Bayesian methods provide a coherent framework for inference, based on the posterior distribution, and for assessing model fit using posterior predictive checks (PPCs). This is done by generating hypothetical replications of the data from the posterior predictive distribution of the model and comparing features of replicated data to observed data. Visual representations of PPCs have been proposed as a generic and flexible tool for model assessment; however, there is little guidance on how to construct PPCs for hierarchical models–particularly outside the scope of HLMs with Gaussian errors. Drawing on a database of repeated measures of curriculum-based K-12 reading and math assessments for at-risk students, we provide concrete examples of how to use PPCs for model checking and comparison. Through these examples, we identify general principles for building PPCs in the context of hierarchical modeling.

Bayesian Treed Regression for Estimating Heterogeneous Trajectories of Test Scores in Large-scale Educational Data

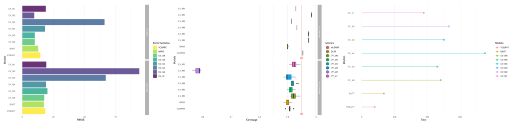
Thursday, 18th July - 09:45: Bayesian Stacking, Trees, and Predictive Modeling (NB C) - Oral

Ms. Mingya Huang (University of Wisconsin - Madison), Prof. Sameer Deshpande (University of Wisconsin - Madison)

In social and behavioral science research, the effects of time-invariant covariates are often time-varying. For instance, the effect of parents' education on children's academic outcomes can evolve over time. To model time-varying effects, parametric models like repeated measure ANOVA and hierarchical linear model (HLM) are commonly deployed. While these highly parametric methods are interpretable, their predictive capacities have not been rigorously compared to those of more flexible nonparametric approaches. We report the results of systematic simulation studies and empirical analyses comparing traditional parametric models to Bayesian nonparametric models based on Bayesian Additive Regression Trees (BART). Our findings indicate that BART-based alternatives can (1) obtain better out-of-sample prediction considering the individual variability over time; (2) provide a reasonable amount of interpretability and uncertainty quantification; and (3) do not incur much more computational burden. This work contributes to the existing social science literature by demonstrating that Bayesian treed regression methods are viable and feasible alternatives to conventional HLM in terms of the prediction of time-varying effects.



Read hall.png



Math hall.png

Variable Selection with Missing Data Using Bayesian Additive Regression Trees

Thursday, 18th July - 10:00: Bayesian Stacking, Trees, and Predictive Modeling (NB C) - Oral

Prof. Sierra Bainter (The University of Miami)

Methods for variable selection—identifying a subset of important predictors for a given outcome from a set of candidate predictor variables—are motivated by a variety of research questions in psychology. Modern variable selection methods to perform principled variable selection and regularization are available within both Bayesian and classical frameworks. Currently available variable selection methods require complete data for the entire set of candidate predictor variables. Suitable procedures for performing variable selection with missing data are needed, however standard methods for handling missing data in the social sciences are model based. An outstanding challenge is how to account for the uncertainty in missing values along with model uncertainty (i.e. which predictors to include in the model). In this talk, we evaluate Bayesian Additive Regression Trees (BART) for variable selection with missing data. BART is a nonlinear ensemble of trees method based on the Bayesian probability model. Importantly, BART incorporates missingness into the predictive model. We present results from a simulation study evaluating the effects of different missingness mechanisms and patterns of missingness on BART results. We also compare the performance of different BART importance score metrics.

How many plausible values?

Thursday, 18th July - 09:00: Topics in Missing Data Analysis (NB D) - Oral

Dr. Paul Jewsbury (Educational Testing Service), Eugenio Gonzalez (Educational Testing Service), Yue Jia (Educational Testing Service), Daniel McCaffrey (Educational Testing Service)

Plausible values (PVs) are imputed values for latent variables (Mislevy, 1991). While prominently used for large-scale assessments (LSAs), the number of PVs required for analysis is unresolved and inconsistent in practice. NAEP generates 20 PVs (Rogers et al., 2020), PISA and PIAAC generate 10 (OECD, 2014, 2016), while TIMSS, PIRLS, IELS and NAPLAN generate 5 (ACARA, 2023; Martin et al., 2020; OECD, 2021b; von Davier et al., 2023). Some researchers use all PVs supplied by the LSA while others use only one (Arikan et al., 2020; Tat et al., 2019). Prior studies on the number of PVs required have reached contradictory and non-definite conclusions (Bibby, 2020; Chung, 2016; Luo & Dimitrov, 2018).

We show analytically and via simulation that the number of PVs used determines the amount of Monte Carlo error on point estimates and standard errors as a function of the between-imputation variance. We derive expressions to determine the number of PVs required to reach a given level of precision when the between-imputation variance is known or precisely estimated. Finally, we analyze real data from a LSA to provide guidelines supported by theory, simulation and real data on the number of PVs to generate and analyze.

Planned Missing Data Designs for Addressing Measurement Reactivity

Thursday, 18th July - 09:15: Topics in Missing Data Analysis (NB D) - Oral

Mark Himmelstein (Georgia Institute of Technology), David Budescu (Fordham University)

Missing data imputation methods are typically aimed at addressing situations where missingness is an incidental, unanticipated, or unwanted nuisance. If data can be assumed to be missing either fully or conditionally at random, many methods are available for imputation without bias, allowing researchers to estimate parameters of theoretical interest. However, one potentially unexplored application of missing data methods is for addressing problems where measurement itself creates a confounding effect, sometimes referred to as measurement reactivity. Measurement reactivity often occurs in pretest-posttest designs, in which researchers want to understand the effect of a treatment or intervention, but the pretest modifies the effect of the intervention on posttest results, limiting generalizability. We present a novel planned missing data research design and a fully Bayesian imputation modeling approach that can allow researchers to omit the pretest, allowing for theoretical inference about treatment effects in the absence of measurement reactivity. We illustrate via simulation and empirical studies based on a common advice taking experimental paradigm. Simulation results demonstrate the structure of the proposed research design and conditions under which the modeling approach is feasible, including required sample size, number of items, and content dimensionality. In our empirical work, we show how the influence of advice on decision making differs as a function of whether a person's independent beliefs were elicited prior to advice exposure. We find clear effects of measurement reactivity and are successfully able to demonstrate how advice taking differs in the presence or absence of measurement reactivity.

Moderated factor analysis with missing data: a multiple imputation approach

Thursday, 18th July - 09:30: Topics in Missing Data Analysis (NB D) - Oral

Dr. Joost Van Ginkel (Leiden University), Dr. Dylan Molenaar (University of Amsterdam)

In moderated factor analysis, the parameters of the traditional linear factor model are a function of an external continuous moderator variable. This approach can be used to, for instance, test for measurement invariance with respect to continuous variables like age and to test for interactions between observed and latent variables. As moderated factor analysis relies on full information maximum likelihood, missing values on the indicator variables of the factor model are relatively unproblematic. However, for cases with missing values on the moderator variable the model likelihood cannot be evaluated. These cases are commonly listwise deleted from the analyses. Consequently, if the moderator variable contains missing data, a method for handling these missing data may be necessary to prevent increased parameter imprecision and/or bias. One such method is multiple imputation. In the current research we investigate bias and coverage of model parameters in a moderated factor model when both the items and the moderator variable contain missing data, and the missing data are handled using multiple imputation. The resulting approach is compared to listwise deletion of the cases with missing data on the moderator variable. Results show that listwise deletion indeed affects the power to detect moderation effects, while for multiple imputation power is only mildly affected. Thus, multiple imputation may be a good way to handling missing data in moderated factor models.

Ising Network Analysis with Missing Data

Thursday, 18th July - 09:45: Topics in Missing Data Analysis (NB D) - Oral

Dr. Siliang Zhang (East China Normal University)

The Ising model has become a popular psychometric model for analyzing item response data. The statistical inference of the Ising model is typically carried out via a pseudo-likelihood, as the standard likelihood approach suffers from a high computational cost when there are many variables (i.e., items). Unfortunately, the presence of missing values can hinder the use of pseudo-likelihood, and a listwise deletion approach for missing data treatment may introduce a substantial bias into the estimation and sometimes yield misleading interpretations. This paper proposes a conditional Bayesian framework for Ising network analysis with missing data, which integrates a pseudo-likelihood approach with iterative data imputation. An asymptotic theory is established for the method. Furthermore, a computationally efficient {P{ó}lya}-Gamma data augmentation procedure is proposed to streamline the sampling of model parameters. The method's performance is shown through simulations and a real-world application to data on major depressive and generalized anxiety disorders from the National Epidemiological Survey on Alcohol and Related Conditions (NESARC).

Validation of an SEM multiverse path test

Thursday, 18th July - 09:00: Topics in Causal Inference (RB 209) - Oral

Ronald Flores (University of Missouri), Prof. Edgar Merkle (University of Missouri)

Many competing theories, each operationalized by SEM, can attempt to explain and estimate the same causal effect. This multiplicity, coupled with the realization that every SEM belongs to a family of equal-fitting SEMs, contributes to a “multiverse” of ways in which to explain and estimate an effect. For the current project, we attempt to validate a test statistic that can assess estimation differences for the same causal effect across sets of competing SEMs. This serves as a sensitivity analysis for researchers to judge whether structural model differences significantly influence a targeted causal effect estimate. Specifically, in cases where an effect estimate does not vary across competing models, the differing structures of the models in which this effect is situated could be ignored, suggesting robust effect estimation. Conversely, if an effect estimate does vary, it could qualify results. We will first give the underlying theoretical details of our SEM multiverse path test and preliminary performance results. Next, we report new performance results using simulated data sets with multiple fitted models that either correctly or incorrectly estimate the same causal effect. We will then close with some discussion of applications and future steps.

Addressing clustered data structures in causal mediation analysis

Thursday, 18th July - 09:15: Topics in Causal Inference (RB 209) - Oral

Hanna Kim (University of Wisconsin-Madison), Prof. Jee-Seon Kim (University of Wisconsin-Madison)

Mediation effects, once deemed identifiable, can be estimated using diverse methods. Parametric closed-form estimators offer clarity in understanding natural direct and indirect effects as they integrate coefficient estimates from regression models. However, their validity hinges on correct model specifications and inclusion of pertinent variables. This study explores the impact of clustering on model accuracy, focusing on the Head Start program's influence on children's vocabulary development. Specifically, children may attend early Head Start programs at age three and learn receptive vocabulary skills by the end of their pre-K year through their reenrollment in regular Head Start programs at age four. Considering that children are clustered into neighborhoods with access to specific Head Start centers, we compare closed-form NDE and NIE estimators under three scenarios: assuming no clustering as in conventional causal mediation analyses, employing cluster-robust standard error estimation, and incorporating fixed effects of clusters. We assess the performance of each of the estimators across study designs with varying degrees and types of clustering through an extensive simulation study. Our findings underscore the importance of adapting analytic methods to reflect clustering in study designs, enhancing the validity of inferences drawn from causal mediation analyses. In sum, clustering effects should be accounted for in parametric causal mediation analyses to mitigate bias and improve result accuracy. By aligning analytic methods with the real-world complexities of clustering, researchers can achieve more robust and valid insights into the causal mechanisms of psychological processes, such as early childhood vocabulary development through early childhood education programs.

An overarching framework for examining diverse forms of heterogeneous causal effects

Thursday, 18th July - 09:30: Topics in Causal Inference (RB 209) - Oral

Prof. Jee-Seon Kim (University of Wisconsin - Madison), Xiangyi Liao (University of Wisconsin - Madison), Mr. Graham Buhrman (University of Wisconsin - Madison), Prof. Wen Wei Loh (Emory University)

Understanding and addressing heterogeneous treatment effects holds significant theoretical and practical importance across many disciplines. This has led to increasing efforts to identify and estimate treatment heterogeneity in various contexts. In this study, we propose an overarching framework for investigating heterogeneous treatment effects. We integrate methodological advances toward examining both observed and unobserved heterogeneity. Our approach encompasses a range of parametric, semi-parametric, and nonparametric methods, such as latent variable modeling, mixture modeling, multilevel modeling, and machine learning techniques. Through real data analyses and realistic semi-synthetic simulation studies, we demonstrate the benefits of collectively incorporating different methodologies to uncover various forms of heterogeneity. Importantly, we explain how to mitigate the risk of making misleading conclusions about treatment effectiveness. The proposed framework provides an accessible platform that enhances the understanding of heterogeneous treatment effects and offers valuable insights and tools for researchers and practitioners seeking to devise targeted interventions and treatment strategies.

Evaluating RKHS mapping effects on functional structural equation models

Thursday, 18th July - 09:45: Topics in Causal Inference (RB 209) - Oral

Dr. Michio Yamamoto (Osaka University / RIKEN AIP), Dr. Yoshikazu Terada (Osaka University / RIKEN AIP)

In recent research, the methodology for examining causal relationships among multivariate functional data has seen significant advancement. This study introduces a structural equation model tailored for functional data, leveraging linear operators to articulate function-on-function regression. In multivariate analysis for functional data, it is often assumed that the functional data are elements of a reproducing kernel Hilbert space (RKHS). If, in reality, the data exist in a broader function space, the analysis is conducted on the data after mapping them onto an RKHS. Our theoretical evaluation reveals that pre-mapping to the RKHS in our proposed model can mitigate the conditions required for the smoothness of variables, ensuring the well-definedness of regression operators. We further illustrate the utility of our model and theoretical findings through numerical examples, highlighting the relaxation of smoothness conditions and the potential for broader applicability in functional data analysis.

Multilevel Regression Discontinuity Analysis with Measurement Models

Thursday, 18th July - 10:00: Topics in Causal Inference (RB 209) - Oral

Youngjin Han (University of Maryland), Muwon Kwon (University of Maryland), Youjin Sung (University of Maryland), Dr. Yang Liu (University of Maryland), Dr. Ji Seung Yang (University of Maryland)

A regression discontinuity (RD) design where a treatment is assigned based on a continuous running variable (RV) and its cutoff allows researchers to estimate a local average treatment effect (ATE) without a randomized control. In this study, we extend the conventional RD analysis by accommodating two common characteristics of data in social sciences: 1) involvement of latent constructs, and 2) multilevel data structure. Consequently, our method enhances the practical utility of RD analysis. By integrating multilevel measurement models into the context of RD analysis, we demonstrate two distinctive benefits of the proposed latent variable modeling approach to RD analysis; quantification of heterogeneous ATEs due to the different levels of latent running variable and generalization of the local ATE away from the cut-off on the observed running variable. In this presentation, we first discuss the definition of causal estimands, model specification, and interpretation of heterogeneity and generalizability of the local ATE. Then the results of Monte Carlo simulation study and empirical data analysis are reported to assess the model estimation and to highlight the utility of the proposed approach in practice. The proposed approach aims at promoting applied research to gather causal evidence for interventions, examine the heterogeneity and generalizability of treatment effect across sites and individuals, and make evidence-based adjustment to treatment administration.

A Deep Learning approach for a more efficient estimation of Rank Preserving Models.

Thursday, 18th July - 09:00: Machine Learning and Deep Learning (RB 210) - Oral

Mr. Roberto Faleh (University of Tübingen), Prof. Holger Brandt (University of Tübingen)

Mediation analysis is an essential method in the empirical sciences, notably in medicine and social sciences, where mediator variables are often critical to determining treatment success in advance. So far, the majority of approaches have been developed under the assumption of no-unmeasured confounders or sequential ignorability which requires that all confounders be measured and incorporated into the model. This assumption can be difficult to satisfy in practice, and violations can result in biased and misleading causal effect estimates. The Rank Preserving Model (RPM; Ten Have et al., 2007) is one of the methods developed for relaxing the assumption of no unmeasured confounders. However, models with the RPM assumption have often not been widely employed due to low power and inefficiency.

We propose a novel Deep Neural Network solution inspired by the TARnet architecture (Shalit et al„2017), augmented with an additional computational graph and custom risk functions. Our proposed methodology aims to mitigate the power and efficiency limitations inherent in traditional RPM frameworks. Using a simulation study we can show that this implementation not only enhances the statistical power and efficiency of RPM but also preserves its semiparametric characteristics of the model, thereby allowing explainable methodological approaches and results.

Modelling computer-based formative assessment data with graph neural networks

Thursday, 18th July - 09:15: Machine Learning and Deep Learning (RB 210) - Oral

Dr. Benjamín Garzón (University of Zurich), Mr. Lisi Qarkaxhija (Julius-Maximilians-Universität Würzburg), Dr. Vincenzo Perri (Julius-Maximilians-Universität Würzburg), Prof. Ingo Scholtes (Julius-Maximilians-Universität Würzburg), Prof. Martin J. Tomasik (University of Zurich)

Computer-based formative feedback (CBFA) systems are software tools for data collection and classroom performance evaluation that enable providing feedback and support instructional decisions. We model a dataset obtained from the MINDSTEPS CBFA system, which serves a population of thousands of students, covering four school subjects and containing millions of responses in a large and highly sparse student-item response matrix. A natural representation of the data is a bipartite graph in which nodes stand for students or items and an edge between a student and an item represents a response of the student to that item (Figure 1A). To predict unobserved edge labels (correct/incorrect responses) we resort to a graph neural network (GNN), a machine learning method recently developed for graph-structured data. The specific model we use consists of an encoder module with two graph convolutional layers followed by a decoder (classifier; Figure 1B). Nodes and edges are represented as embeddings in a multidimensional space, and the model can also incorporate student (e.g., gender), item (e.g., competence domain) and response (e.g., age) features.

After fitting the GNN, the learned item embeddings recover properties of the school curriculum, such as item difficulty and the structure of school subjects and competences (Figure 2). We propose that the model parameters can be used to inform curriculum development in a data-driven manner, and conclude by briefly discussing the advantages and disadvantages of the proposed approach with respect to more established alternatives.

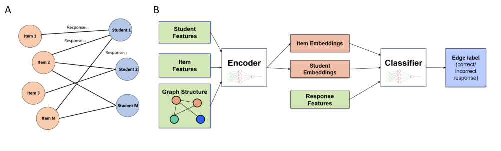


Figure 1. A) Bipartite graph, where nodes are students or items and edges indicate that a student has responded to a certain item. B) Simplified model architecture.

Figure1.png

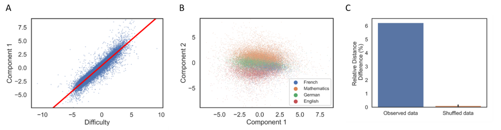


Figure 2. A) The first component of a principal component analysis of the item embeddings was strongly correlated with item difficulty. B) Item embeddings captured the structure of school subjects, such that item embeddings of different school subjects tended to be further apart than those from the same subject (relative distance difference was positive and significant, C).

Figure2.png

Residual Permutation Tests for Feature Importance in Machine Learning

Thursday, 18th July - 09:30: Machine Learning and Deep Learning (RB 210) - Oral

Dr. Po-Hsien Huang (National Chengchi University)

Psychological research has traditionally relied on linear models for testing scientific hypotheses. However, the emergence of Machine Learning (ML) algorithms has opened up new possibilities for exploring variable relationships beyond linear constraints. To interpret the outcomes of these “black-box” algorithms, several interpretation tools for feature importance have been developed. Nonetheless, such tools are largely descriptive and do not facilitate statistical inference. In response to this gap, our study introduces two versions of the Residual Permutation Test (RPTs), designed to assess the significance of a target feature in predicting the label. The first variant, RPT on y (RPT-Y), permutes the residuals of the label conditioned on features other than the target feature. The second variant, RPT on x (RPT-X), permutes the residuals of the target feature conditioned on the other features. Our simulation study demonstrates that RPT-X effectively maintains empirical Type I error rates within acceptable bounds and demonstrates appreciable power in both regression and classification tasks, suggesting its utility for hypothesis testing in ML applications.

Parameterized Material Recommendation with Dual-goal Adaptive Learning System

Thursday, 18th July - 09:45: Machine Learning and Deep Learning (RB 210) - Oral

Dr. Tongxin Zhang (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China), Prof. Tao Xin (Beijing Normal University)

Adaptive learning system (ALS) recommends appropriate learning materials based on learners' characteristics. Current researches of ALS focus on single-goal especially the knowledge growth, barely take high-order thinking abilities into account, e.g., information processing ability (IPA). As Kalyuga et al. (2003) notes, instructing materials (e.g., texts) are more suitable for learners with low IPA while non-instructing materials (e.g., experiments) are more appropriate for experts with high IPA, i.e., the Expertise Reversal Effect. And IPA grows once learners with low IPA acquire knowledge successfully from non-instructing materials by random searching. Thus, material recommendation is a trade-off between short-term adaption and long-term ability growth. In this study, we implemented an ALS facilitated by reinforcement learning (RL) to optimize the dual-goal of knowledge growth and IPA acquirement in the simulated scenarios. We parameterized the materials based on their instructional/non-instructional characteristics to make them available to RL and applied DINA model to generate measurement error of learners' characteristics. The results of simulated studies show that: Firstly, the effect of dual-goal ALS is well since that most learners can fulfill knowledge learning tasks (i.e., S_31 in Fig.1) and all learners can achieve high IPA after a 12-step adaptive learning process. Secondly, the recommendation strategy of materials is interpretable and reasonable. RL tends to recommend non-instructional materials at the early stage of learning process to improve IPA, later RL will recommend materials adaptive to learners with different level of IPA (See Fig.2). This study shows the potential to achieve more complex and multiple learning goals in ALS.

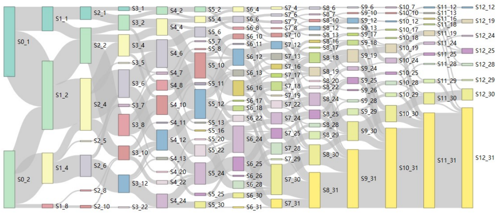


Figure 1. transfer flow of learners knowledge states within 12 learning steps.jpg

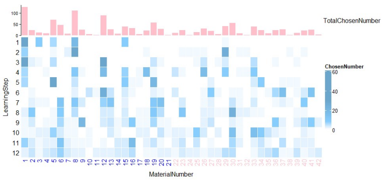


Figure 2. learning material recommendation frequency within 12 learning steps.jpg

Bayesian Dynaimic Mediation Model

Thursday, 18th July - 09:00: Bayesian Factor Analysis and SEM (RB 211) - Oral

Mr. CHEN Qijin (Department of Psychology, Sun Yat-sen University), Prof. PAN Junhao (Department of Psychology, Sun Yat-sen University)

Most existing mediation analysis is stationary assuming that the mediation effect is time-invariant, neglecting the fact that most human psychological processes and behaviors are dynamic. The current research proposes Bayesian time-varying structural equation models, which combine time-varying mediation models and confirmatory factor analysis, to investigate dynamic mediation effects. Considering both measurement error and regression residuals, simulation studies show that the proposed method works well and reflects the true nature of mediation process. Besides, we use empirical data as an example to illustrate how our models work in real research.

Bayesian SEM fit indices: now for categorical indicators

Thursday, 18th July - 09:15: Bayesian Factor Analysis and SEM (RB 211) - Oral

Dr. Mauricio Garnier-Villarreal (Vrije Universiteit Amstedam), Dr. Terrence Jorgensen (University of Amsterdam), Prof. Edgar Merkle (University of Missouri)

Model fit is a crucial step of SEM research. As Bayesian methods become more available, we have developed some common methods from frequentist SEM to BSEM, such as approximate fit indices for BSEM models with continuous indicators. As other developments happened we need to catch up and update methods, we have now user friendly BSEM models with categorical indicators. And have found that the methods for fit indices for continuous indicators do not directly translate to categorical indicators. In this project we present a proof of concept of a method to estimate posterior distributions of chi-square statistics for BSEM models with categorical indicators, and then this are used to estimate posterior distributions of approximate fit indices such as CFI, RMSEA and gamma-hat. The discrepancy measure estimated across posterior draws requires to be corrected by a factor of the effective number of parameters (similar to the previous indices), but the chi-square statistics has to be estimated in a different matter accounting for the categorical structure of the data. We will present preliminary results of a small simulation to evaluate the accuracy of this method to estimate posterior distributions of fit indices. This simulation will seek to test it across model type, complexity, misfit level, and sample size.

Bayesian fit measures in approximate measurement invariance in cross-cultural research

Thursday, 18th July - 09:30: Bayesian Factor Analysis and SEM (RB 211) - Oral

Dr. Chunhua Cao (University of Alabama), Dr. Xinya Liang (University of Arkansas), Ms. Lijin Zhang (Stanford University), Ms. Yangmeng Xu (The University of Alabama)

In extensive cross-cultural studies involving many countries, achieving exact full measurement invariance poses a challenge. Exploring approximate measurement invariance (AMI) (Muthén & Asparouhov, 2013) becomes pivotal for a comprehensive understanding of measurement invariance status in large-scale data. The AMI method relaxes the strict measurement invariance assumptions by employing a zero-mean, small-variance prior to accommodate minor variation in measurement parameters across groups. This methodological flexibility has led to the increased adoption of AMI in diverse applications. Despite its growing popularity, the effectiveness of Bayesian fit indices, specifically the Bayesian Comparative Fit Index (BCFI) and the Bayesian Root Mean Square Error of Approximation (BRMSEA), in the context of invariance model selection has not been extensively evaluated. Our study addresses this gap through a comprehensive simulation analysis, assessing the performance of various Bayesian fit measures, such as the posterior predictive p-value (PPP), BCFI, BTLI, BRMSEA, BIC, and DIC, in the context of AMI within cross-cultural research with many groups. The design factors include the sample size, number of groups, degree of AMI, and model size. The priors on parameter differences between groups were varied to evaluate the sensitivity of fit indices to prior distributions. Implications to many group comparisons in cross-culture settings will be discussed.

Bayesian Confirmatory Factor Analysis with Missing Values

Thursday, 18th July - 09:45: Bayesian Factor Analysis and SEM (RB 211) - Oral

Prof. Christian Aßmann (Leibniz Institute for Educational Trajectories, Bamberg), Dr. Markus Pape (University of Bochum)

Confirmatory item factor analysis is common for analysing relations between educational assessments, corresponding competence factors, and possible determinants thereof. Bayesian estimation is routinely discussed in the literature to allow for efficient handling of multidimensional latent variables and arbitrary discrimination structures involving zero and equality restrictions. However, missing values inevitably occur in survey data, either by design or due to item non-response. To cope with these missing values, a Bayesian
estimation routine of a multivariate normal-ogive IRT model is presented dealing with missing values in a stringent manner. Data augmentation established for handling of latent factors is thereby extended towards missing values in conditioning variables. The properties of the suggested approach are investigated by means of a simulation study using model setups with typical discrimination structures. Results suggests
that efficiency gains are possible with respect to parameter expectation estimates. Empirical illustration is provided in the context of modelling competence development using data on the starting cohort of fifth grader surveyed in the National Educational Panel Study.

Bayesian joint modal estimation for sparse item factor analysis

Thursday, 18th July - 10:00: Bayesian Factor Analysis and SEM (RB 211) - Oral

Dr. Kensuke Okada (The University of Tokyo), Mr. Keiichiro Hijikata (The University of Tokyo), Mr. Motonori Oka (The London School of Economics and Political Science)

Owing to the widespread adoption of digital- and technology-enhanced assessment tools, there is an increasing demand for item factor analysis in situations in which the numbers of both respondents and items are large. However, existing Bayesian approaches have limitations regarding scalability to large datasets. In addition, it is desirable to obtain sparse factor loadings to enhance interpretability. To address these issues, we propose a Bayesian joint modal estimation method using a sparsity-promoting prior. Specifically, we consider the joint likelihood for the respondent and item parameters with a Laplace prior distribution for the factor loadings; normal priors are set for the other parameters. The proposed estimation algorithm alternates the optimization of the respondent and item parameters until convergence is achieved. Because the update rules for the parameter values are independent for each respondent and item, the proposed algorithm can be implemented in parallel, allowing fast and efficient optimization. The hyperparameter that controls the degree of sparsity can be determined using a cross-validation approach. We conducted a simulation study to compare the proposed method with the Markov chain Monte Carlo estimation, and the results confirmed that the proposed method provided accurate and fast estimation. Moreover, we applied the proposed method to the Synthetic Aperture Personality Assessment dataset comprising numerous respondents and items. The results obtained using the cross-validation-based hyperparameters revealed an interpretable factor structure corresponding to the Big Five personality factors, indicating the efficiency and applicability of the proposed method.

The Investigation of Mediating Processes as a Measurement Challenge

Thursday, 18th July - 10:45: Keynote Address (Vencovského aula) - Keynote Address

Prof. David MacKinnon (Arizona State University)

The investigation of mediating mechanisms is a measurement challenge. Progress is made as the mediating variable process is more precisely measured. In this view, researchers conduct studies to distill or uncover the most important facet of a construct that transmits the causal effect of an independent variable to a dependent variable. Examples in chemistry and genetics illustrate how improved measurement of unobservable processes helped researchers identify critical mediating mechanisms. In psychology, a variety of methods are used to measure mediating processes including self report and biological measures. A strength of the application in psychology is that humans can report information about their feelings, states, and behaviors. A weakness is that humans are susceptible to response tendencies and other factors that increase error and bias in measurements. The goal of the presentation is to discuss measurement approaches for investigating mediating variables and to review the measurement literature on how quality of measurement affects mediation analysis. Topics include the influence of confounding variables, use of alternative measurement models, and measurement overlap on mediation analysis. I discuss future methodological and substantive directions in the measurement of mediating processes including programs of research that include measurement of a mediating construct as a central goal.

Item Response Theory and Health Outcomes: Novel Applications and Methods

Thursday, 18th July - 13:15: Symposium: Item Response Theory and Health Outcomes: Novel Applications and Methods (Vencovského aula) - Symposium Overview

Dr. Edward Ip (Wake Forest University School of Medicine), Dr. Brooke Magnus (Boston College), Dr. Shelley Liu (Icahn School of Medicine at Mount Sinai), Dr. Leah Feuerstahler (Fordham University)

Evaluating health outcomes and biomedical data with standard psychometric tools such as item response theory often involves nonstandard and varied item response formats (e.g., skewed continuous and zero-inflated variables), heterogeneous target populations, multiple measurement attributes within an item, and unknown dimensionality. Thus, there is a growing need for psychometric innovation in analyzing health and biomedical data, ranging from patient-reported outcomes to biomarker data. However, assessing health outcomes also provides opportunities for psychometric innovation. Some recent innovations include incorporating auxiliary information from substantive research about the studied phenomenon, developing novel applications of item response theory to environmental health and exposomics data, and measuring unipolar constructs exhibiting excess zero responses such as depression. This symposium brings together both methodological and applied researchers who use item response theory models to address challenges in the measurement and scaling of health outcomes and biomedical data.
Presenters for this symposium will be:
Dr. Edward H. Ip, Wake Forest University School of Medicine
Dr. Brooke Magnus, Boston College
Dr. Shelley H. Liu, Icahn School of Medicine at Mount Sinai
Dr. Leah Feuerstahler, Fordham University

Latent Variable Modeling for Unipolar Constructs in Health Sciences

Thursday, 18th July - 13:15: Symposium: Item Response Theory and Health Outcomes: Novel Applications and Methods (Vencovského aula) - Symposia

Dr. Edward Ip (Wake Forest University School of Medicine), Dr. Shyh-Huei Chen (Wake Forest University School of Medicine)

Patient Reported Outcome (PRO) data have gained increasing prominence in both medical research and FDA-related regulatory spaces. Inherited from the traditional measurement paradigm, PROs are structured to measure bipolar traits – i.e., traits that have meaning at both ends of the scale. However, in the medical context, certain constructs such as depression and alcoholism manifest as unipolar traits, implying that the trait is only meaningful at one end of the distribution but not the other. For example, low score signifies the absence of a quality (e.g., not alcoholic) and not a relatively low score when compared to others with that quality (i.e., less alcoholic). In other words, the two groups are qualitatively different and putting them on the same scale may not be appropriate. Also there often exist items that are signal potentially alarming condition (e.g., suicidal ideation in a depression inventory) that are highly discriminating than other items (e.g., feeling blue). Furthermore, methods such as inflated zero may not work well for model unipolar traits because non zero low score, in the example of depression, can also indicate the absence of the condition. Not unlike standard setting in the context of educational assessment, one approach is to determine thresholds to indicate different categories of the condition. The threshold approach however, often requires elaborate expert input and consensus. To address these challenges, the presentation will explore several procedures, including IRT, latent class/transition model, and IRT for partially ordered set responses to construct measurement models that are intended for unipolar traits.

The Application of IRT to Traumatic Brain Injury Research: Some Promise and Challenges

Thursday, 18th July - 13:15: Symposium: Item Response Theory and Health Outcomes: Novel Applications and Methods (Vencovského aula) - Symposia

Dr. Brooke Magnus (Boston College)

For nearly 50 years, the Glasgow Outcome Scale–Extended (GOSE) and its predecessor, the Glasgow Outcome Scale (GOS), have been considered the gold standard measures of global outcome after traumatic brain injury (TBI). The GOSE, which is currently the only outcome measure that has been accepted by the U.S. Food and Drug Administration for use in TBI research supporting New Drug Application approvals, places individuals with TBI into one of eight broad levels of injury-related disability. While this classification structure is easy to understand, its simplicity is not always optimal, particularly when more granular assessment of individuals' injury recovery is desired. The GOSE is customarily assessed using a multi-question interview that contains richer information than is reflected in the standard, ordinal GOSE score. Over the past decade, researchers have begun to consider the role that item response theory (IRT) can play in achieving more precise measurement of TBI outcome, with the goal of yielding more informative classification. This talk will provide an overview of the potential benefits of applying IRT to TBI outcome measures – primarily, the ability to obtain scores that show greater sensitivity to change, thus making them better suited for use in clinical trials. It will also discuss some of the associated challenges, including extreme local dependence, differential item functioning across timepoints, and potential mode effects.

Mixture Item Response Theory to Quantify Cumulative Environmental Exposure Burden

Thursday, 18th July - 13:15: Symposium: Item Response Theory and Health Outcomes: Novel Applications and Methods (Vencovského aula) - Symposia

Dr. Shelley Liu (Icahn School of Medicine at Mount Sinai), Dr. Leah Feuerstahler (Fordham University), Ms. Yitong Chen (Icahn School of Medicine at Mount Sinai), Dr. Joseph Braun (Brown University), Dr. Jessie Buckley (University of North Carolina - Chapel Hill)

Quantifying a person's cumulative exposure burden to environmental toxicants is important for risk assessment. However, different people may be exposed to different sets of environmental toxicants due to heterogeneity in exposure sources and patterns. We used mixture item response theory to estimate a person's total exposure burden to per- and polyfluoroalkyl substances (PFAS), dubbed toxic forever chemicals, while accounting for the fact that different people have different diets and behaviors that may expose them to different sets of PFAS chemicals. This ensures that PFAS burden scores can be equitably compared across population subgroups. We applied our methods to PFAS biomonitoring data from the National Health and Nutrition Examination Survey (NHANES) 2013-2018, where we found that Asian Americans have significantly higher PFAS burden compared with non-Hispanic Whites, but this disparity was masked when using summed PFAS concentrations as the exposure metric. Our work suggests that risk assessment may want to consider a summary exposure metric for environmental toxicants that accounts for exposure heterogeneity, so that the summary metric used is fair and informative for all people.

An Explanatory IRT Model for Toxicity-Weighted Environmental Exposure Burden Scores

Thursday, 18th July - 13:15: Symposium: Item Response Theory and Health Outcomes: Novel Applications and Methods (Vencovského aula) - Symposia

Dr. Leah Feuerstahler (Fordham University), Ms. Yitong Chen (Icahn School of Medicine at Mount Sinai), Dr. Shelley Liu (Icahn School of Medicine at Mount Sinai)

Ubiquitous environmental contaminants, such as per- and polyfluoroalkyl substances (PFASs) can be detected in the blood of >98% of the United States population, and are linked to poor liver health and liver cancer in humans. Because PFASs are a large chemical class, regulatory and clinical biomonitoring guidelines recommend quantifying a person's cumulative exposure to PFASs. Recently, our team has successfully used item response theory (IRT) modeling to derive a PFAS exposure burden score that quantifies a person's total latent exposure burden to PFASs, with each PFAS biomarker considered an "item". However, not all PFASs are equally toxic, motivating the need for a toxicity-informed model using health-endpoint specific relative potency factors (RPFs) derived from toxicokinetic models. Here, we describe how we use these RPFs for liver effects in conjunction with explanatory IRT to develop a PFAS burden score specific to liver toxicity. Specifically, we used transformations of the RPFs as fixed discrimination parameters and polychotomized data from the 2017-2018 United States National Health and Nutrition Examination Survey (NHANES) to estimate threshold parameters for the graded response model. We compared the distributions of PFAS burden sum scores, IRT-derived burden scores, and liver toxicity-informed IRT-derived burden scores across body mass index (BMI) categories. Results show that in males, participants who were overweight had significantly higher burden on all measures compared to normal-weight participants, though no significant differences were found in females. IRT burden scores significantly predicted some indicators of liver functioning after covariate adjustment, while results using sum scores were null.

A Psychometric Analysis on Associations between Coding Concepts and Item Difficulty

Thursday, 18th July - 13:15: Problems in Cognitive Diagnostic Modeling (RB 101) - Oral

CHEN Li (Educational Testing Service), Mo Zhang (Educational Testing Service), Hongwen Guo (Educational Testing Service), Xiang Liu (Educational Testing Service), Amy Ko (University of Washington), Min Li (University of Washington)

Computer science has become increasingly popular among college students, with many beginning their studies by learning a programming language like Python in introductory courses. These courses often assess students' programming skills through tasks that require writing computer programs, evaluated using test cases. These tasks cover various programming concepts, including integer manipulation, looping, and function recursion, each with its own difficulty level. Despite the prevalence of such assessments, there is a lack of psychometric analysis in the literature. This presentation aims to fill this gap by exploring the application of psychometric analysis and modeling to a computer programming assessment dataset. In addition to assessing the correctness of the final programs, the analysis also considers process data such as time spent and number of attempts. By analyzing these data, we aim to uncover potential associations between different programming concepts and difficulty levels. Moreover, we propose fitting a cognitive diagnostic model (CDM) with process data as covariates to classify students into different skill mastery patterns. This modeling approach can provide valuable insights for instructors and researchers in understanding how students learn and master programming concepts.

Evaluating methods for assessing model fit in diagnostic classification models

Thursday, 18th July - 13:30: Problems in Cognitive Diagnostic Modeling (RB 101) - Oral

Jake Thompson (University of Kansas)

In recent years, Diagnostic Classification Models (DCMs) have received more attention from the psychometric community. Much of this research has focused on developing new models, estimation algorithms, and model fit indices. However, relatively little attention has been paid to the application of DCMs and how existing research findings translate to practice. This is especially true for model fit, where numerous absolute and relative fit indices have been proposed, but never compared to each other in a way that can inform decisions made by applied practitioners. The current study aims to begin filling this gap by comparing how different measures of absolute and relative model fit for DCMs perform under realistic test designs. In this study, the Loglinear Cognitive Diagnostic Model (LCDM) and Deterministic Inputs, Noisy "And" gate (DINA) models are used to generate data and estimate a DCMs in a factorial design, allowing us to evaluate the effectiveness of model fit indices for different test designs under conditions where true model fit is known. We then apply the findings to an empirical example from an operational K-12 assessment in the United States that uses DCMs to for reporting results to teachers, parents, and students. Results of the study are used to make recommendations for selecting model fit indices when utilizing DCMs in practice. In addition, we introduce the R package measr, which can be used to estimate a variety of DCMs and calculate all model fit methods described in the talk.

A general diagnostic modeling framework for forced-choice items

Thursday, 18th July - 13:45: Problems in Cognitive Diagnostic Modeling (RB 101) - Oral

Dr. Pablo Nájera (Universidad Pontificia Comillas), Dr. Rodrigo Schames Kreitchmann (National University of Distance Education), Scarlett Escudero (Universidad Autónoma de Madrid), Dr. Francisco J. Abad (Universidad Autónoma de Madrid), Prof. Jimmy de la Torre (The University of Hong Kong), Dr. Miguel A. Sorrel (Universidad Autónoma de Madrid)

Over the past two decades, the diagnostic classification modeling (DCM) framework has undergone numerous advances both in terms of methodological developments and domains of applications. One particular direction in which this modeling approach can be expanded is through the incorporation of forced-choice (FC) assessments, which have been shown to mitigate the problems related to response biases in high-stakes contexts. Recently, Huang (2023) achieved an important milestone in this direction by proposing a restricted DCM specifically tailored for FC items. The present work aims to build upon this research by integrating FC assessments into an established DCM framework, namely the G-DINA model framework. The proposed approach offers several advantages, which include a more straightforward conceptualization of FC item parameters, enhanced flexibility in item response functions and attribute structures, a more computationally efficient estimation using existing R packages, and the possibility of adopting the suite of analyses developed for traditional DCM (e.g., classification accuracy estimation, model fit evaluation) to FC assessments. To ensure proper interpretation of the results, important considerations for attribute identification when implementing these analyses using the 'GDINA' and 'cdmTools' R packages are discussed. Finally, the viability of the proposed general framework is evaluated by means of a simulation study. Overall, this work extends the G-DINA framework to FC assessments, providing a robust foundation for conducting comprehensive diagnostic assessments of non-cognitive traits.

Correcting validity coefficients with aggregated data in cognitive diagnosis models

Thursday, 18th July - 14:00: Problems in Cognitive Diagnostic Modeling (RB 101) - Oral

Dr. Rodrigo Schames Kreitchmann (National University of Distance Education), Dr. Pablo Nájera (Universidad Pontificia Comillas), Dr. Miguel A. Sorrel (Universidad Autónoma de Madrid), Dr. Francisco J. Abad (Universidad Autónoma de Madrid)

The primary goal of measurement is to extend the understanding of scores, allowing interpretations beyond the measurement context itself. As a result, external validity plays a central role in psychological and educational assessments. Within Cognitive Diagnosis Models (CDMs), it may be desired, for instance, to inspect how the mastery of different levels of learning (e.g., as in Bloom's taxonomy) may influence educational outcomes. However, classifications from CDMs are seldom error-free, causing attenuation in the estimated relationships with external variables. Some strategies, including a one-step approach estimating measurement and structural models simultaneously, or a corrected three-step approach accounting for measurement error in class estimates, have been proposed to correct this underestimation.

This study introduces a new method to address attenuation of validity coefficients caused by classification errors. The proposed approach relies exclusively on aggregated data, using descriptive statistics of the external variables within the estimated classes, along with classification accuracy. A key advantage is its suitability for meta-analyses, as it eliminates the necessity for the actual response data. Additionally, beyond the correction of validity coefficients, it also provides corrected descriptive statistics within the latent classes.

To illustrate the properties of the corrected validity estimators, a simulation study is presented, offering insights into bias and consistency of the proposed correction approach. Five factors were manipulated in the study, including the true validity coefficient, generated and estimated CDM, test length, sample size, and item quality. The results support the correction approach as an unbiased and consistent estimator of the true validity coefficients.

A generalized Q-matrix estimation method for nonparametric cognitive diagnosis

Thursday, 18th July - 14:15: Problems in Cognitive Diagnostic Modeling (RB 101) - Oral

Mrs. Jia Li (Beijing Normal University), Dr. Ping Chen (Beijing Normal University)

The Q-matrix is a crucial component of cognitive diagnostic assessment, which builds a bridge between the items and attributes. In practice, the Q-matrix is generally developed by researchers and domain experts and may contain misspecifications. To address the issue, researchers have proposed a variety of data-driven methods over the past decade. Compared with the parametric Q-matrix estimation methods, the nonparametric methods do not require a complex and time-consuming parameter estimation process and can also achieve high estimation accuracy under small sample conditions. However, there are currently few nonparametric Q-matrix estimation methods, and they cannot be directly used in general models.

Inspired by ideas from signal detection theory, a decision criterion based on the ideal responses (DC-IR) method that is not model-constrained has been proposed in this paper with a view to providing methodological support for nonparametric Q-matrix estimation in general models. Two simulation studies were conducted not only to evaluate the performance of the DC-IR method but also to compare it with the existing residual sum of squares (RSS; Chiu, 2013) and generalized nonparametric classification (GNPC; Chiu et al., 2018) methods.

The results showed that the new method significantly outperformed the GNPC method in the G-DINA model, and the estimation accuracy in the DINA model was comparable to that of the RSS method. Future research could consider generalizing the new method to test scenarios with polytomous scoring or polytomous attributes, as well as exploring the use of machine learning methods for nonparametric Q-matrix estimation with small samples.

Covariate-Adjusted Generalized Factor Analysis with Application to Testing Fairness

Thursday, 18th July - 13:15: Topics in Factor Analysis (NB A) - Oral

Jing Ouyang (University of Michigan), Chengyu Cui (University of Michigan), Dr. Kean Ming Tan (University of Michigan), Dr. Gongjun Xu (University of Michigan)

In the era of data explosion, statisticians have been developing interpretable and computationally efficient statistical methods to measure latent factors (e.g. skills, abilities, and personalities) using large-scale assessment data. In addition to understanding the latent information, the covariate effect on responses controlling for latent factors is also of great scientific interest and has wide applications, such as evaluating the fairness of educational testing, where the covariate effect reflects whether a test question is biased toward certain individual characteristics (e.g. gender and race) taking into account their latent abilities. However, the large sample size, substantial covariate dimension, and great test length pose great challenges to developing efficient methods and drawing valid inferences. Moreover, to accommodate the commonly encountered discrete type of responses, nonlinear factor models are often assumed, bringing in further complexity to the problem. To address these challenges, we consider a covariate-adjusted generalized factor model and develop novel and interpretable conditions to address the identifiability issue. Based on the identifiability conditions, we propose a joint maximum likelihood estimation method and establish estimation consistency and asymptotic normality results for the covariate effects under a practical yet challenging asymptotic regime. Furthermore, we derive estimation and inference results for latent factors and the factor loadings. We illustrate the performance of this method through extensive numerical studies and an application to a large-scale educational assessment, the Programme for International Student Assessment (PISA).

A factor analytic analogue for summed score EAPs

Thursday, 18th July - 13:30: Topics in Factor Analysis (NB A) - Oral

Dr. Alexis Georgeson (Arizona State University)

Summed score EAPs are a pragmatic scoring approach developed within Item Response Theory. In contrast to response pattern EAPs in which every response pattern has a different corresponding EAP, summed score EAPs are computed so that each sum score has one corresponding EAP. The advantage of this scoring approach is for applications in which greater precision is desired, but software for computing response pattern EAPs is not available or practical. To our knowledge, there is not yet an existing analogue of summed scored EAPs within factor analysis. The challenge is that with truly continuous item responses and total scores, it is not as straightforward to create the summed score EAPs. Using both simulated and real data, this presentation will investigate the feasibility of such a scoring method, called summed score factor scores, and show results highlighting different possible approaches for computing them, and their consequences.

Comparing item selection in regularized factor analysis and network models

Thursday, 18th July - 13:45: Topics in Factor Analysis (NB A) - Oral

Ms. Jiaying Chen (University of Arkansas), Dr. Xinya Liang (University of Arkansas), Dr. Jihong Zhang (University of Arkansas)

This study seeks to explore and compare the efficacy of regularized psychometric network models (Epskamp et al., 2017) and regularized factor analysis models (e.g., Jacobucci, et al., 2016) in analyzing complex data structures, focusing on model fit and item selection. Psychometric network and factor analysis provide different interpretations but utilize comparable techniques. Network modeling analyzes partial correlations between variables to identify network structure, key nodes, and pathways for intervention, while factor analysis uncovers underlying factor structures that explain observed variable covariances. Both methods can incorporate regularization techniques such as Lasso to aid in simplifying models by penalizing less important parameters, thereby enhancing model estimation and preventing potential overfitting. A unique aspect of this research involves examining whether instruments developed through factor analytic framework are appropriate for subsequent network analysis. As unidimensional structure remains a common choice for instrument validation (e.g., He et al., 2021), our study targets this model structure. In addition to comparing the model fit of regularized models from both frameworks, this study will evaluate the item selection processes specific to each framework - network analysis employs a two-step approach involving redundancy analysis followed by regularization to eliminate weak nodes, in contrast to factor analysis, which utilizes a direct application of regularization on model parameters for item selection. The aim is to understand if network analysis identifies items of similar importance to those identified by factor analysis models, thus contributing valuable insights into the comparative strengths and applications of the two frameworks in complex data analysis.

Multilevel factor mixture modeling: A simulation study

Thursday, 18th July - 14:00: Topics in Factor Analysis (NB A) - Oral

Dr. Eunsook Kim (University of South Florida), Dr. Chunhua Cao (University of Alabama), Dr. Yan Wang (University of Massachusetts Lowell)

When data are multilevel (e.g., individuals nested within clusters), heterogeneity can arise from multiple sources, such as variations across individuals and the clusters to which they belong. Multilevel factor mixture modeling (MLFMM) incorporates both latent factors and latent classes at multiple levels and allows researchers to identify unobserved heterogenous groups at both individual and cluster levels. The benefits of MLFMM such as for multilevel interventions (cluster-based intervention as well as individual-based intervention) have been illustrated (Cao et al., 2022), but its applications are rare in practice. The systematic review of FMM (Kim et al., 2023) showed the majority of FMM applications using multilevel data either ignored nesting (i.e., single-level FMM) or accounted for nesting using adjusted standard errors (i.e., design-based FMM) instead of building MLFMM, but the appropriateness of these approaches has not been investigated. Furthermore, the efficacy of MLFMM to detect heterogeneity at different levels is unknown. To address these gaps in the literature, we conduct a Monte Carlo simulation study (1) to investigate the impact of using either single-level FMM or design-based FMM and (2) to examine the efficacy of MLFMM under various multilevel research conditions including multilevel factor structure, effect size, indicator intraclass correlation (ICC), factor ICC, number of clusters, cluster size, and class proportion conditions. The simulation outcomes include class enumeration, classification accuracy, and parameter recovery. We provide practical guidelines for an optimal approach to modeling multilevel latent classes along with multilevel latent factors given research conditions (e.g., factor structure, ICC and sample size).

Pseudo-factor analysis of embedding similarity matrices

Thursday, 18th July - 14:15: Topics in Factor Analysis (NB A) - Oral

Nigel Guenole (Goldsmiths, University of London), Andrew Samo (Bowling Green State University), Tianjun Sun (Kansas State University)

We show that pre-knowledge of item discrimination is obtainable as the cosine similarity between transformer embeddings of items and their construct definitions, which we refer to as 'pseudo-discrimination' (i.e., discrimination calculated without response data). First, we discuss results from a pre-print by Guenole, Samo, & Sun (2024) that studied pseudo-discrimination for 2 personality measures and 4 sentence transformer models. Pseudo-discrimination values showed internal convergent and discriminant validity for pseudo-discrimination, with pseudo-discriminations being high for items on their target scales and low for items on non-target scales. We next show that for these same measure and transformer combinations, cosine similarities between trait definitions and items predicted empirical factor loadings for the majority of scales studied (4 out of 6 scales for a DSM-5 based measure, 7 out of 10 scales for BFAS-100). We suggest discriminative power is conferred on pseudo-discrimination because pseudo and empirical discrimination both provide alternative perspectives on the same part (items) to whole (trait) relationship, and that despite being different quantities calculated in different ways from different data, both represent the fidelity, or 'belonginess' of items to constructs. We then generalise the approach to show that we can fit factor analysis models to cosine similarity matrices of item embeddings and see evidence of expected theoretical structures using the IPIP NEO and IPIP HEXACO inventories. We will present results of comparisons between pseudo and empirical factor structures and discuss implications of pseudo-discrimination for test design.

References.

Guenole, N., Samo, A., Sun, T. (2024). Pseudo-discrimination parameters from language embeddings. https://osf.io/preprints/psyarxiv/9a4qx

Recent advances in (non-)Bayesian informative hypothesis evaluation

Thursday, 18th July - 13:15: Symposium: Recent advances in (non-)Bayesian informative hypothesis evaluation (NB B) - Symposium Overview

Dr. Rebecca Kuiper (Utrecht University)

Hypothesis evaluation is a key feature of research in the behavioral, social, and biomedical sciences. For example, researchers might be interested in whether a new medicine A works better (e.g., leads to more happiness) than and old medicine B, which still works better than placebo (in an ANOVA model: $\mu_A > \mu_B > \mu_{Placebo}$); or: whether the number of children is a stronger predictor of happiness than income and age (in a regression model with standardized parameters: $\beta_{NoC} > \{\beta_{Inc}, \beta_{age}\}$); or: whether the cross-lagged effect of stress to anxiety, of rumination to stress, and of rumination to anxiety are higher than the cross-lagged counterparts (in a random-intercept cross-lagged panel model: $\varphi_{SA} > \varphi_{AS}, \varphi_{RS} > \varphi_{SR}, \varphi_{RA} > \varphi_{AR}$). Evidence for such **informative, theory-based hypotheses** cannot be provided by null hypothesis testing but can be obtained via (information-theoretical and Bayesian) model selection.

In this symposium, we will first provide a conceptual and technical introduction to the topic: We will demonstrate why informative hypotheses are crucial for progress in social sciences, especially with regards to theory-building and addressing the replication crisis (by using informative hypotheses in pre-registration). Second, we will discuss several specific applications to substantive research, namely psychotherapy research, media psychology, psychometrics, and cognitive psychology. In the symposium, we will showcase both the Bayesian and information-theoretical model-selection approaches to the evaluation of informative hypotheses. Third, we will discuss the need for a comparison frame and provide guidelines for the interpretation of information-theoretical model selection results (using study-specific benchmarks).

Applying the Informative Hypothesis Approach to Media Effects

Thursday, 18th July - 13:15: Symposium: Recent advances in (non-)Bayesian informative hypothesis evaluation (NB B) - Symposia

Ms. Nikol Kvardova (Interdisciplinary Research Team on Internet and Society, Masaryk University)

On account of the ongoing replication and theory crisis in psychology, it is necessary to implement effective tools for theory-building in psychology research. The field of media psychology, which centers on studying how media affect people in both the short-term and the long-term, has not been an exception to this urgent need for theory-driven research efforts. As an example, the social media effects on well-being have been of significant interest of public and scholarly attention lately. Yet, the results of these efforts have so far been inconclusive and ambiguous, unable to further inform the public or scholarly debate. One way to bring this research area forward could be by making our theories more precise and stronger, allowing us to draw informative, testable predictions. By strengthening the theories and guiding researchers to make their expectations more explicit and specific, using informative hypotheses might also help in navigating the replication crisis. In this presentation, I will present an example of applying GORIC(A) to test the hypotheses drawn from the Sociocultural Theory of Body Image (Thompson et al., 1999), exploring the longitudinal impact of social media on body dissatisfaction. I will demonstrate the process of formulating informative hypotheses and evaluating them with GORIC(A). Furthermore, I will discuss the implications for theory-building and addressing the replication crisis in media psychology, the lessons from which can be applied to other substantive fields.

Theory-based hypothesis evaluation using information criteria

Thursday, 18th July - 13:15: Symposium: Recent advances in (non-)Bayesian informative hypothesis evaluation (NB B) - Symposia

Dr. Rebecca Kuiper (Utrecht University)

Hypothesis evaluation is a key feature of research in the behavioral, social, and biomedical sciences. For example, researchers might be interested in whether a new medicine A works better (e.g., leads to more happiness) than and old medicine B, which still works better than placebo (in an ANOVA model: $\mu_A > \mu_B > \mu_{Placebo}$); or: whether the number of children is a stronger predictor of happiness than income and age (in a regression model with standardized parameters: $\beta_{NoC} > \{\beta_{Inc}, \beta_{age}\}$); or: whether the cross-lagged effect of stress to anxiety, of rumination to stress, and of rumination to anxiety are higher than the cross-lagged counterparts (in a random-intercept cross-lagged panel model: $\varphi_{SA} > \varphi_{AS}$, $\varphi_{RS} > \varphi_{SR}$, $\varphi_{RA} > \varphi_{AR}$). Evidence for such **informative, theory-based hypotheses** cannot be provided by null hypothesis testing but can be obtained via (information-theoretical and Bayesian) model selection.

In my presentation, I will focus on model selection using information criteria. I will introduce the AIC-type criterion GORIC and its approximation: the GORICA. I will discuss how the GORIC(A) results can be interpreted and how they quantify the support for the theory-based hypothesis (as opposed to null hypothesis testing). I will address two cases: i) the case when there is one study and ii) the meta-analytic case when there are multiple studies from which one would like to aggregate the results (where, for this method, the multiple studies can be both exact and conceptual replications).

Bayesian Evaluation of N=1 Studies

Thursday, 18th July - 13:15: Symposium: Recent advances in (non-)Bayesian informative hypothesis evaluation (NB B) - Symposia

Prof. herbert hoijtink (Utrecht University)

The Bayes factor and corresponding posterior model probabilities can be used to evaluate a set of competing hypotheses. The set can contain a null hypothesis, one or more informative hypotheses (e.g., b1 > 0, b2 >0 & b3 >0, where the b's denote regression coefficients), and the complement of these hypotheses. First of all, Bayes factor and posterior model probabilities will introduced. Subsequently, two N=1 studies will be introduced. The first concerns tracking three outcome variables for a patient that is receiving trauma therapy for a period of thirteen weeks. The hypotheses of interest are H0: b1 = 0, b2 = 0 & b3 = 0, that is there is no effect of the therapy on the outcome measures; H1: b1 < 0, b2 < 0 & b3 > 0, that is, two outcome measures decrease and one increases as a result of the therapy; and, the complement of the union of both hypotheses, that is, something else is going on. It will be shown that thirteen measurements of three outcome variables is enough to evaluate these hypotheses. The second example concerns tracking a family (parents and two children) during the 52 weeks in which they receive counseling. In each week the inter-family experience level of violence is recorded for each family member. Hypotheses regarding the development of "experience level of violence" will be formulated and evaluated. The presentation is concluded with a short discussion.

Making Nomological Networks Confirmatory using GORICA

Thursday, 18th July - 13:15: Symposium: Recent advances in (non-)Bayesian informative hypothesis evaluation (NB B) - Symposia

Mr. Petr Palisek (Psychology Research Institute, Faculty of Social Sciences, Masaryk University), Dr. Rebecca Kuiper (Utrecht University)

Researchers often use nomological networks to argue for the validity of their measure. They observe a set of correlations between their measure and other variables and then proceed to judge whether these associations are what the relevant theory would imply. Such a process is often problematic because, lacking clear, study-specific interpretational guidelines, the researcher can choose to accept multitudes of nomological networks by post hoc reasoning that these were, in fact, expected all along. In this presentation, we show how informative hypothesis evaluation using GORICA can be utilised to avoid this problem by (1) allowing for clear and strict representation of the expected nomological network while (2) offering an intuitive way to quantify support for these expectations given the observed data. Connecting nomological networks with the concept of informative hypotheses evaluation brings this validation tool closer to its originally intended, highly theory-driven purpose.

Benchmarks and cut-off norms: Necessary evils or questionable science inducers?

Thursday, 18th July - 13:15: Symposium: Recent advances in (non-)Bayesian informative hypothesis evaluation (NB B) - Symposia

Lars de vreugd (University Medical Center Utrecht)

Comparisons are fundamental in almost everything we do (Goldstone et al., 2010), provide the basis for making judgements (Laming, 2004), and underpins decision-making and drawing inferences (Gentner et al., 2001). Comparisons can be essential in Learning Analytics, internal feedback generation, and in statistics as well - e.g. the Cohen's *d* effect size benchmark, or $p < .05$ as cut-off value. The latter, used in Null Hypothesis Significance Testing, contributed to the 'file-drawer problem' and the replication crisis (Wasserstein et al., 2019). Suggested remedies are reporting Confidence Intervals, *s*-values, or *p*-values as continuous descriptive statistics, to avoid dichotomous reject/do-not-reject decision making.

Other approaches avoid *p*-values altogether, by determining informative hypotheses (e.g., m1 > m2 > m3) a-priori and quantifying support by using, for instance, the Bayes factor or GORIC weights. For the Bayes factor, threshold values for "positive" (>3) and "strong" (>20) evidence have been suggested (Kass & Raferty, 1995), but this can lead to "3" becoming the Bayes equivalent of "p<.05". However, if comparisons are fundamental and integral to cognitive processes, how does one interpret statistical output without comparison frames? Furthermore, can we reasonably expect non-statisticians to incorporate aforementioned advanced statistical practices into their research?

In this presentation, I will discuss possibilities in balancing the need for comparison to interpret statistical output but at the same time avoiding oversimplified decision-making. I will also discuss a study of GORIC interpretation guideline development, challenges applied researchers encountered when using them, and the latest guideline developments to support researchers when interpreting GORIC output.

Emotions under control? The relationship between emotionality and cognitive control reverse if event-unpleasantness is increased within individuals

Thursday, 18th July - 13:15: Symposium: Recent advances in (non-)Bayesian informative hypothesis evaluation (NB B) - Symposia

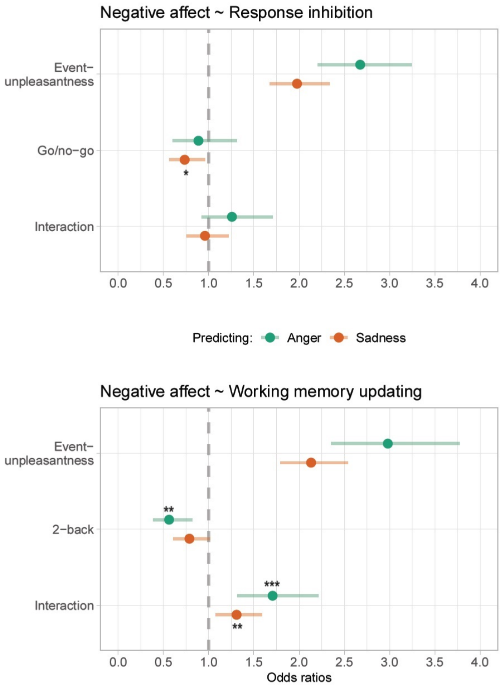
Mr. Levente Rónai (Institute of Psychology, ELTE, Eötvös Loránd University; Institute of Psychology, University of Szeged), Ms. Flóra Hann (Institute of Psychology, ELTE, Eötvös Loránd University), Prof. Szabolcs Kéri (National Institute of Mental Health, Neurology and Neurosurgery – Nyírő Gyula Hospital; Department of Cognitive Science, Budapest University of Technology and Economics), Prof. Ulrich Ettinger (Department of Psychology, University of Bonn), Dr. Bertalan Polner (Institute of Psychology, ELTE, Eötvös Loránd University; Donders Institute for Brain, Cognition and Behaviour, Radboud University)

Impaired cognitive control is associated with maladaptive emotion regulation across individuals. Yet, little is known about whether this relationship holds within individuals. Here, we tested the assumption that momentary within-person increases in working memory updating and response inhibition performance predict heightened emotional reactivity in everyday life.

Participants from the general population repeatedly (8 two-hourly prompts daily) performed short 2-back and Go-Nogo tasks using their own devices in daily life. Affective states and event-unpleasantness were assessed, and emotional reactivity was modeled as their association. We fitted cumulative link mixed models in two overlapping samples: a Go-Nogo (N = 161; N[obs.] = 2494) and a 2-back dataset (N = 158; N[obs.] = 2641).

Our analyses revealed that when individuals' momentary working memory updating was better than their average, they demonstrated higher negative emotional reactivity. However, better working memory performance predicted decreased negative affect if event-unpleasantness was at individuals' average. Better Go/no-go performance predicted lower negative emotionality but not reactivity. These results were also confirmed by the informative hypothesis testing approach using 'restriktor' and 'benchmarks' R packages.

These results may question the idea that improved cognitive control is universally related to adaptive emotion regulation. While it seems improbable that emotional reactivity enhances working memory, future research should clarify the direction of causality.



Tweetbrat.jpg

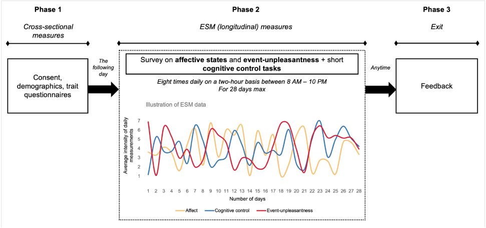

Phase 1
Cross-sectional measures
Phase 2
ESM (longitudinal) measures
Phase 3
Exit
Consent, demographics, trait questionnaires
The following day
Survey on affective states and event-unpleasantness + short cognitive control tasks
Eight times daily on a two-hour basis between 8 AM – 10 PM
For 28 days max
Anytime
Feedback
Illustration of ESM data
Average intensity of daily measurements
1
2
3
4
5
6
7
Number of days
Affect
Cognitive control
Event-unpleasantness

Fig 1 cog cont design 1 page-0001.jpg

Extending the time-varying temporal network to allow changing contemporaneous connections

Thursday, 18th July - 13:15: Topics in Intensive Longitudinal Data Analysis (NB C) - Oral

Simran Johal (University of California, Davis), Emilio Ferrer (University of California, Davis)

When applied to intensive longitudinal data, psychometric network models produce two networks: a temporal network and a contemporaneous network. The former describes the dynamic relations between variables over time; the latter describes the within-occasion associations between variables after accounting for temporal relations. Although methods have been proposed to allow the relations within the temporal network to vary over time (Bringmann et al., 2018; Haslbeck et al., 2021), the contemporaneous network has been assumed to remain time-invariant (Rast, 2023).

We propose a new method that extends one of the current estimation approaches for time-varying temporal networks to allow the contemporaneous network to also vary over time. Using the same generalized additive modeling framework used to estimate the time-varying temporal network, we estimate the contemporaneous network using multiple regression models applied to the residuals of the temporal network. This framework allows the regression coefficients quantifying the contemporaneous associations to change over time in a gradual manner. We examine the performance of the proposed method through a Monte Carlo simulation study under multiple data conditions, including the number of variables in the network model, the number of timepoints, the pattern of change in the contemporaneous network over time, the approach used during network estimation to set coefficients to zero (e.g., significance thresholding), and the number of basis functions used in the estimation procedure as factors. Finally, we demonstrate the implementation of our proposed approach with an application using intensive longitudinal data.

Individual Versus Group-Level Analysis Techniques for Intensive Longitudinal Data: A Primer

Thursday, 18th July - 13:30: Topics in Intensive Longitudinal Data Analysis (NB C) - Oral

Conor Lacey (University of North Carolina at Chapel Hill), Dr. Kathleen Gates (University of North Carolina at Chapel Hill), Jennifer Traver (University of North Carolina at Chapel Hill), Hannah Lewis (University of North Carolina at Chapel Hill)

Increasing popularity in Intensive longitudinal analysis (ILA) has resulted in a plethora of within-individual methodological research practices. While the development of these methods has proven invaluable for understanding the intricacies of individual-level processes over time, little guidance has been provided on which methods are appropriate for various research goals. A particular problem that burdens the field is the choice between individual-level or group-level analysis techniques, for the decision of which method to use can lead to substantively different conclusions. To provide insight on this issue, this paper explores the strengths and weakness of autoregressive time-series methods and multi-level modeling methods. Specifically, we explore how autoregressive (AR), vector autoregressive (VAR), vector autoregressive (mlVAR), and multi-level models (MLM) differ under varied sample sizes, time-points, and number of predictors as well as under conditions of effects heterogeneity. Based on the results of our simulations, we propose a variety of recommendations including the use of group-level analyses when working with fewer timepoints and the use of individual-level analyses when working with small sample size. The ultimate goal of this paper is meant to highlight the advantages and disadvantages of utilizing certain ILA methods to better inform the research community who utilize intensive-longitudinal data.

Investigation of detrending methods for intensive longitudinal dyadic data analysis

Thursday, 18th July - 13:45: Topics in Intensive Longitudinal Data Analysis (NB C) - Oral

Yue Xiao (East China Normal University), Hongyun Liu (Beijing Normal University)

The intensive longitudinal data using dyads as basic units, i.e., the intensive longitudinal dyadic data (ILDD), can be used for investigating the dynamic process of interpersonal interaction within dyads and the differences of such processes between dyads. For ILDD in which each person is paired with only one other person, Savord et al. (2022) extended the popular Actor-Partner Interdependence Model (APIM) in the framework of Dynamic Structural Equation Modeling (DSEM) to describe the dynamic relationship of the same or different variables between two members in each dyad. However, the specification of residual covariance between dyad members in this model does not fully explain the interdependence between two members. More importantly, the model does not take the issue of detrending into account. As the DSEM framework involves time series modeling, nonstationary time series without detrending may bias the parameter estimation, further distorting the relationships between variables. At the same time, it remains unclear whether the detrending methods within DSEM framework for independent individuals are applicable to the dyadic data. In this study, we reconstruct a more general model for the ILDD based on Savord et al.'s model and integrate two detrending approaches within DSEM framework by including an additional covariate (i.e., time factor). Through a simulation study, we examined and compared the performance of two integrated models in terms of estimation accuracy and power of the parameters of primary interest for intensive longitudinal dyadic studies. Afterward, recommendations are provided for applied researchers based on the simulation study results.

Estimating actor and partner effects from intensive longitudinal dyadic data with large amounts of missing values: Estimation performance of the longitudinal actor-partner interdependence model and possible alternatives Abstract

Thursday, 18th July - 14:00: Topics in Intensive Longitudinal Data Analysis (NB C) - Oral

Yuanyuan Ji (KU Leuven), Jordan Revol (KU Leuven), Anna Schouten (KU Leuven), Marieke Schreuder (KU Leuven), Prof. Eva Ceulemans (KU Leuven)

Researchers interested in dyadic processes increasingly make use of intensive longitudinal study designs. The longitudinal actor-partner interdependence model (L-APIM) provides an appealing modeling approach for such data. However, intensive longitudinal data are almost always incomplete, due to non-compliance, and the use of conditional questions. These missing data issues become more prominent in dyadic data, because dyadic partners often do not miss the same measurement occasion. Moreover, partners might disagree about features that trigger conditional questions. Large amounts of missing data challenge the estimation performance of the L-APIM. Specifically, we found that non-convergence occurred when applying the L-APIM to pre-existing dyadic diary data with a large amount of missing values. Using a simulation study, we therefore systematically examined the performance of the L-APIM in intensive longitudinal dyadic data with missing-data issues. In line with the findings for our illustrative data, we found that non-convergence often occurred in conditions with a small sample size, while the estimation of actor and partner effects remained relatively good in case analyses did converge. Additionally, in light of potential convergence failures with the L-APIM, we aimed to provide insight in alternative models for investigating dyadic processes. We proposed nine alternative models and evaluated their performance on simulated and empirical data. Overall, when the L-APIM fails to converge, we recommend fitting multiple alternative models to check the robustness of the results.

Using Structural Equation Models to Analyze Round-Robin Data from Social Networks

Thursday, 18th July - 13:15: Symposium: Using Structural Equation Models to Analyze Round-Robin Data from Social Networks (NB D) - Symposium Overview

Dr. Terrence Jorgensen (University of Amsterdam)

Study social phenomena from an interpersonal perspective is enabled by round-robin designs, in which each member of a group provides data about every other member—e.g., each student in a classroom indicates how much they like each other student, or each nuclear-family member indicates how secure their relationship is with each other family member. This complex pattern of interdependence among dyadic observations (Y_{ij}: a variable Y measured about person i responding to or interacting with person j) has a social-network structure, which requires sophisticated analytical models to account for interdependence. The linear social relations model (SRM: $Y_{ij} = \mu + P_i + T_j + R_{ij}$) decomposes such interpersonal perceptions into random effects associated with perceivers (P_i), their targets (T_j), and relationship-specific nuances captured by dyad-level residuals (R_{ij}). Sampling several round-robin groups (e.g., families, classrooms) also enables modeling of group-level variance via a random intercept μ_g rather than constant mean μ. A multivariate SRM can estimate correlations among multiple round-robin variables, but fitting theoretical models to explain those relationships requires a larger modeling framework, such as structural equation modeling (SEM). This symposium highlights how to analyze SRM data with SEM via open-source software. The first pair of presentations shows how to use the traditional SEM-framework for modeling the SRM and complex psychometric extensions. The second pair of presentations compare and evaluate 1- and 2-stage maximum-likelihood estimation methods for analyzing multivariate-SRM data via the social-relations SEM (SR-SEM). All presentations and software examples are provided on the Open Science Framework (OSF): https://osf.io/z8xay

Multiple constructs multiple indicators social relations modes with and without roles

Thursday, 18th July - 13:15: Symposium: Using Structural Equation Models to Analyze Round-Robin Data from Social Networks (NB D) - Symposia

Dr. Fridtjof W. Nussbeck (University of Konstanz), Dr. David Jendryczko (University of Konstanz)

Structural Equation Modeling (SEM) can serve as an overarching framework to model social relations models (SRM). We show how the standard SRM can be extended to model multiple constructs assessed with multiple indicators each. We present the model for non-interchangeable, interchangeable and partly interchangeable members of the round robin group. Furthermore, we present how model fit can be adequately judged disentangling model fit assessment of interchangeability from the assessment of the substantial hypotheses. Advantages, caveats, possible extensions, and limitations in comparison to alternative modeling strategies are discussed. Illustrative examples for the presented analyses can be obtained from the OSF-repository of the symposium.

A correlated traits correlated (methods – 1) multitrait-multimethod model for augmented round-robin data

Thursday, 18th July - 13:15: Symposium: Using Structural Equation Models to Analyze Round-Robin Data from Social Networks (NB D) - Symposia

Dr. David Jendryczko (University of Konstanz), Dr. Fridtjof W. Nussbeck (University of Konstanz)

We show how the traditional structural equation modeling approach for the multiple constructs multiple indicators social relations model can be extended to model a correlated traits correlated (methods - 1) [CTC(M – 1)] multitrait-multimethod model for augmented round-robin data (round-robin data that also include self-reports). We present the variance decomposition as well as consistency and reliability coefficients. Moreover, we explain how to evaluate fit of a CTC(M – 1) model for augmented round-robin data. Results of a simulation study suggest good properties of the full information maximum likelihood estimation of the model, even in cases where the reciprocity-covariance structure of the augmented round-robin design is miss-specified. Implications and limitations are discussed. Illustrative examples for the presented analyses can be obtained from the OSF-repository of the symposium.

Toward a general multivariate framework for social network data: An overview of estimation methods for structural social-relations models

Thursday, 18th July - 13:15: Symposium: Using Structural Equation Models to Analyze Round-Robin Data from Social Networks (NB D) - Symposia

Dr. Terrence Jorgensen (University of Amsterdam)

Models of social-network data must account for interdependencies among dyadic observations (Y_{ij}) within a round-robin group (i.e., each group member i responds about or interacts with each other member j). The social relations model (SRM) is a linear decomposition of (approximately) continuous variables into person-level random effects—perceivers (P_i) and their targets (T_j)—and dyad-level relationship effects (R_{ij}). Univariate-SRM analyses investigate relative contributions of each effect's variance component, as well as correlations among person-level effects (generalized reciprocity) and dyad-level effects (dyadic reciprocity). Univariate SRM has been extended to allow predictors of person- and dyad-level effects, and multivariate SRM enables estimating correlations between (person- and dyad-level components of) multiple round-robin variables. Various ad-hoc methods have been used to parsimoniously explain such relationships via regression or structural equation models (SEM). We review two-stage estimation procedures to estimate person- and dyad-level effects, which are then treated as data in a subsequent SEM. The recently developed social-relations SEM (SR-SEM) uses single-stage maximum-likelihood estimation (MLE) to avoid bias introduced by treating estimates as data, but it lacks the flexibility of two-stage approaches—e.g., to model only person- or dyad-level effects and include person-level covariates. We propose a new two-stage MLE technique using estimated summary statistics as data, which is flexible without sacrificing validity of inferential statistics (or requiring the computational burden of other two-stage solutions) and makes it possible to overcome other remaining limitations (e.g., assuming multivariate normality). We discuss the computation details and implementation in the R package lavaan.srm.

Incorporating level-specific covariates in structural equation models of social-network data

Thursday, 18th July - 13:15: Symposium: Using Structural Equation Models to Analyze Round-Robin Data from Social Networks (NB D) - Symposia

Ms. Aditi Manoj Bhangale (University of Amsterdam), Dr. Terrence Jorgensen (University of Amsterdam)

The social relations model (SRM) is applied to examine multivariate dyadic data (e.g., round-robin data) within social networks. Such data have a unique nesting structure in that dyads are cross-classified within individuals who may be nested within different social networks. The SRM decomposes perceptual measures into multiple individual- and dyad-level components (incoming, outgoing, and relationship effects), the associations among which were previously estimated using linear random-effects models. However, such models cannot estimate complex structural relations between SRM components of different variables, for which one might use a structural equation model (SEM). The social-relations SEM (SR-SEM) combines the SRM and SEM, enabling researchers to test several measurement and structural hypotheses regarding SRM components. However, incorporating level-specific covariates—such as age, gender, or relationship quality—as predictors and outcomes of SRM components remains a challenge. In this study, we propose a two-stage estimation approach to easily incorporate level-specific covariates into SEMs of round-robin variables. Stage 1 of the two-stage estimator is Markov chain Monte Carlo estimation of unrestricted SRM summary statistics. Stage 2 is maximum likelihood estimation of constrained SEMs using the Stage-1 summary statistics of SRM effects as input data. A previous simulation revealed that diffuse priors yield inaccurate Stage-1 estimates, which negatively impacted the accuracy of Stage-2 results. Thus, we investigate different empirical-Bayes priors and compare their impact on Stage-2 accuracy and efficiency of level-specific covariate effect estimates.

Keywords. Social relations model, structural equation model, two-stage estimation, level-specific covariates, prior specification

Detecting Non-Uniform Differential Item Functioning in Latent Variable Interaction Framework

Thursday, 18th July - 13:15: Differentital Item Functioning (RB 209) - Oral

Dr. Heining Cham (Fordham University), Ms. Xinyue Deng (Fordham University), Ms. Hyunjung Lee (Fordham University)

Non-uniform differential item functioning (DIF) can be represented by the interaction effect between the latent variable and groups. This study investigates different statistical procedures to detect non-uniform DIF. First, we investigate the utility of the Homoscedastic Fit Index (HFI; Gerhard, Büchner, Klein, & Schermelleh-Engel, 2017) to detect the omitted interaction terms when non-uniform DIF exists. Second, we compare various methods (i.e., sum scores, factor scores, and latent variable interaction) to estimate the DIF parameters and perform null hypothesis tests. Historically, researchers have used the sum scores of items as a proxy for the latent variable and employed generalized linear modeling to estimate the interaction effect for non-uniform DIF (e.g., Chen & Jin, 2018). Ng and Chen (2020) suggested that factor scores can be an alternative proxy for the latent variable. The past simulation by Woods and Grimm (2011) found that the latent moderated structural equation (LMS) approach, which is one algorithm for estimating the latent variable interaction effect, could result in an inflated Type I error rate when testing for non-uniform DIF. Third, we examine the feasibility of estimating the newer moderation effect size measures by Liu and Yuan (2021) to quantify non-uniform DIF. This study aims to furnish researchers with a thorough comprehension of various methods for identifying non-uniform DIF, as well as to offer practical guidelines regarding the significance of these effect sizes.

Extending the Cluster Approach to Differential Item Functioning in Polytomous Items

Thursday, 18th July - 13:30: Differentital Item Functioning (RB 209) - Oral

Mr. Martijn Schoenmakers (Tilburg University), Dr. Jesper Tijmstra (Tilburg University), Prof. Jeroen Vermunt (Tilburg University), Dr. Maria Bolsinova (Tilburg University)

To objectively compare different groups on any latent trait using tests, the absence of differential item functioning (DIF) is crucial. While the importance of DIF has been well-established in the psychometric literature, the question of how to adequately select DIF-free items is still largely open, with many different approaches being proposed. The fact that the difficulty of an item is not identified from the observations alone may be a reason no widely agreed upon approach to DIF testing has been developed. Recently, DIF tests utilizing the differences between item difficulties across groups, which are identified, were proposed for the Rasch (Bechger & Maris, 2015) and 2-parameter logistic (2PL) models (Pohl et al., 2021). The current paper aims to extend this clustering approach to the polytomous case using the partial credit model (Masters, 1982). To achieve this, the clustering approach to DIF in a polytomous item is split into two steps. First, the distances between the item thresholds within an item are compared across groups using a multivariate Wald test. If these item threshold distances are not found to differ for an item, the item proceeds to the second stage. Here, items are clustered on the differences between a single item threshold across groups. Due to previously establishing the lack of differing item thresholds within an item across groups, this is equivalent to clustering items on differences between all thresholds across groups. Performance of the new approach is assessed using a simulation study and practical recommendations are made.

Advancing Polytomous DIF Detection with the Residual DIF Framework

Thursday, 18th July - 13:45: Differentital Item Functioning (RB 209) - Oral

Dr. Hwanggyu Lim (Inha University), Dr. Jaime Malatesta (Graduate Management Admission Council), Dr. Yongsang Lee (Inha University)

Lim et al. (2022) introduced the residual-based differential item functioning (RDIF) detection framework, comprising RDIF$_R$, RDIF$_S$, and RDIFR$_{RS}$ statistics, each specifically designed to test uniform, nonuniform, and mixed DIF, respectively.

The RDIF framework offers a multitude of advantages, including minimal computational demand, satisfactory power, well-controlled Type I error rates, and the elimination of the need for group-specific item calibrations, IRT equating, or sequential model fitting. Building on this, Lim et al. (2023a) demonstrated the adaptability of the RDIF framework for DIF screening in computerized adaptive tests, while Lim et al. (2023b) further broadened its scope, showing its applicability in detecting DIF across multiple groups.

This study contributes to the RDIF literature by introducing two additional unique approaches suitable for polytomous DIF detection. Computational summaries for the two approaches, for an item with five score categories, is given below:

1. Using aggregated item score data, compute residuals between observed item scores (e.g., 2) and model-predicted scores (e.g., 2.5).
2. Using individual score category data, calculate residuals between one-hot encoding vectors (e.g., [0, 0, 1, 0, 0] for an observed score of 2) and the corresponding probability vectors (e.g., [.1, .2, .4, .25, .05]), representing the probability of endorsing each score category.

Preliminary simulation results suggest that both extended versions of the RDIF framework effectively maintain well-controlled Type I error rates and have sufficient power. Additionally, their notably swifter implementation compared to other IRT-based methods positions them as practical and efficient tools for evaluating DIF in polytomous items.

Assessing differential step functioning for polytomous items with scale purification

Thursday, 18th July - 14:00: Differentital Item Functioning (RB 209) - Oral

Dr. Ya-Hui Su (National Chung Cheng University), Ms. Fu-Mei Liu (National Chung Cheng University)

Differential item functioning (DIF) occurs when examinees with the same ability but from different groups have different probability to answer an item correctly. If the different probability happened to each step of a polytomous item, it is called differential step functioning (DSF). Many DIF studies have been conducted to develop DIF detection methods for polytomous items, but most of the methods focus on the item-level DIF assessment. Therefore, Penfield (2007) extended the Liu-Agresti estimator (LA), which is the item-level DIF assessment, to the simultaneous step-level test (SSL), which is the step-level DSF assessment, for polytomous items. Penfield manipulated only one item might have DIF in the study, and the item was the only studied item. In practice, more than one items might be contaminated DIF in the test, and all items should be considered for DIF detection. Besides, DSF might happened to any step, and DSF might favor different groups with different amount. Thus, this study investigated the efficiency of the DSF detection with these two methods with scale purification under various conditions for polytomous items. The results showed that the Type I error of both methods met model expectations when the overall test effect size was 0. When the overall test effect size was increased, the Type I error of both methods was inflated. The SSL method outperformed the LA method in flagging DSF even when items had unbalanced DSF. Furthermore, applying scale purification could improve the performance of both methods.

Psychometric innovations for adaptive learning systems

Thursday, 18th July - 13:15: Symposium: Psychometric innovations for adaptive learning systems (RB 210) - Symposium Overview

Dr. Maria Bolsinova (Tilburg University)

The move toward personalized learning holds the promise of making tailor-made education available to everyone through development of large-scale online adaptive learning systems (ALS), allowing each learner to maximally realize their learning potential and improving both the learning process and the learning outcomes. These systems provide a large amount of data in terms of responses of students to practice items, which need to be modeled appropriately to optimize feedback, instructions, learning materials and subsequent practice items that are presented to students. In this coordinated session we will discuss methodological and practical challenges that are created by the complexity of these data (large-scale, highly sparse, with intricate missing data patterns) and present psychometric innovations that address them. The first three talks deal with the issue of ability measurement in ALS which is not a trivial due to the adaptive, dynamic and large-scale nature of ALS. The first two presenters (Hanke Vermeiren and Abe Hofman) will discuss methodological advances and practical challenges in the use of the Elo rating system for dynamic updating of student abilities and item difficulties. An alternative system for tracking ability is the focus of the third presentation (Bence Gergely) with the specific interest in improvement of measurement by incorporation of response time data. The last two talks are not concerned with measurement per se, but offer psychometric solutions for related issues, namely modeling individual differences in quitting behavior (Annie Johansson) and diagnostics of errors (Alexander Savi), and deliver important insights for better personalization of online education.

A dynamic k value approach for the Elo rating system

Thursday, 18th July - 13:15: Symposium: Psychometric innovations for adaptive learning systems (RB 210) - Symposia

Hanke Vermeiren (KU Leuven), Dr. Abe Hofman (University of Amsterdam), Dr. Maria Bolsinova (Tilburg University), Prof. Han L. J. van der Maas (University of Amsterdam), Prof. Wim Van den Noortgate (KU Leuven)

In adaptive digital learning environments, it is essential to track learning trajectories. The Elo rating system, known for its computational simplicity, is frequently employed for this purpose. Current Elo-based systems cannot handle rapid changes in ability or are unable to balance accuracy and speed when updating player and item ratings. Changes in Elo ratings depend on the sensitivity parameter k. Using fixed k values necessitates a trade-off. Larger values facilitate the tracking of evolving ability levels but introduce greater rating volatility. Smaller values yield more stable estimates, but are slower to reflect actual ability levels. Existing modifications of the Elo rating system, which diminish K as the number of responses increases, are inadequate in scenarios characterized by considerable ability fluctuation, a common occurrence in digital learning environments. To address this challenge, we introduce a novel approach for dynamically adjusting k values in response to observed trends in rating changes. This method increases k during noticeable upward or downward shifts in ratings and reduces it otherwise. We present a computationally efficient implementation of this idea and validate its superiority over existing k adjustment strategies through simulation studies. Additionally, we describe the implementation of this adaptive k model in a widely used digital learning platform, Math Garden, which leverages both accuracy and response time in its assessments.

Curious interactions between learners and Elo rating systems

Thursday, 18th July - 13:15: Symposium: Psychometric innovations for adaptive learning systems (RB 210) - Symposia

Dr. Abe Hofman (University of Amsterdam)

In this talk, I will share lessons learned from designing and maintaining a large adaptive learning platform (Math Garden, Language Sea). I will highlight several cases that show an interesting interaction between the Elo rating system and (unexpected) player behaviour. In these cases, self-reinforcing feedback loops between ratings and behaviour can have negative effects on the measurement properties and eventually break the system. These results show the importance of monitoring our adaptive learning platform.

Tracking changing abilities in adaptive learning system using response times

Thursday, 18th July - 13:15: Symposium: Psychometric innovations for adaptive learning systems (RB 210) - Symposia

Bence Gergely (Eötvös Lóránd University, Doctorate School of Psychology), Dr. Maria Bolsinova (Tilburg University)

Adaptive Learning Systems (ALS) dynamically adjust the level of practice based on the student's abilities, providing optimally engaging content. To enable this, one needs a continuously updated and reliable measure of the changing student abilities. Rating systems, such as the Elo rating system and the Urnings algorithm provide a computationally inexpensive method to achieve this. However, these algorithms base their estimation on response accuracy, which is often not sufficient to reliably track individual-level ability changes.

To increase the precision of the measurement, we developed methods to incorporate response times (RT) into the Urnings algorithm. First, we combined the accuracy and the RT into a single variable using a scoring rule. Then we created a stream of conditionally independent dichotomous variables, by using dyadic expansions on the continuous score. Using these new discrete variables we defined two ways to sequentially update the parameter estimates of the Urnings algorithm, by either deploying the original algorithm for each dichotomous variable or updating a single rating multiple times.

In a simulation study, we demonstrated that both modified methods yielded faster convergence compared to the original Urnings algorithm if the step size of the updates were the same. We further showed that the modified algorithms increase the prediction accuracy even if a limited number of dyadic expansions were used based on the reanalysis of an existing ALS.

The new methods provide better predictions and converge faster to their (known) limiting distribution while providing unbiased estimates and known standard errors.

Modeling individual differences in error-induced quitting.

Thursday, 18th July - 13:15: Symposium: Psychometric innovations for adaptive learning systems (RB 210) - Symposia

Ms. Annie Johansson (University of Amsterdam), Dr. Abe Hofman (University of Amsterdam), Prof. Han L. J. van der Maas (University of Amsterdam), Dr. Alexander Savi (University of Amsterdam)

Learning systems that utilize adaptive algorithms can deliver and enhance education in a large-scale, individualized, and accessible manner. Despite this potential, many of these systems suffer low retention rates. Further, individual characteristics such as motivation, grit, and self-regulation are all necessary for successful engagement in learning but are also traits that vary between learners and add a layer of complexity in how algorithms should adapt to the learner. In this talk, I will discuss how (big) educational data can be leveraged to examine how engagement-related factors vary across individuals. Specifically, I will present a study conducted in a large-scale computer adaptive system that models to what extent sequential errors (making several mistakes in a row) predict quitting. Results show strong average effects of quitting after making one, two, three or more than three incorrect responses in the system. However, modeling individual-specific effects in this relationship showed that individuals vary considerably in the extent that sequential errors induce quitting, with some students displaying an increased tendency to quit, some unaffected, and some persisting more following errors. These individual differences also proved to be stable across different learning games. Our results corroborate the theoretical notion that students differ in their tolerance to failure and pinpoint a need to individualize how computer-adaptive systems intervene after errors. I will discuss the challenges in modeling individual differences in cognitive- and motivational variables and highlight important future steps to move away from a one-size-fits-all approach to education.

The future of educational measurement: Integrating AI and psychometrics

Thursday, 18th July - 14:45: Cito Special Symposium: The future of educational measurement: Integrating AI and psychometrics (Vencovského aula) - Symposium Overview

Dr. Joost Kruis (Cito Institute for Educational Measurement)

The rapid development of artificial intelligence (AI) technologies is reshaping the landscape of educational measurement and assessment. In this symposium, we explore the intersections of AI with traditional psychometric approaches, with the aim of addressing the complexities and challenges of using AI in educational contexts. Our symposium aims to highlight innovative strategies, ethical considerations, and methodological advances in integrating AI into educational measurement.

The first presentation will address the socio-technical challenges of AI in education, highlighting the need for holistic validation strategies that incorporate both quantitative and qualitative methods to ensure the trustworthiness of educational technologies. The second presentation addresses the pressing issue of scale harmonisation across research consortia, demonstrating the application of machine learning and test-equating methods to facilitate meaningful comparison of ADHD scores. Moving forward, the third presentation evaluates the potential natural language processing models, in automating the scoring of open-ended responses, thereby increasing the efficiency and objectivity of this type of assessment. In an attempt to bridge the past and the future, the final presentation revisits the historical contributions of psychometrics to data science, reminding us that many contemporary AI applications have their roots in psychometric practices.

Through these discussions, our symposium aims to chart a course for the judicious and innovative integration of AI into educational measurement, acknowledging both its promise and its pitfalls, while emphasising the enduring relevance of psychometric principles.

Safeguarding the Human Element in Educational Measurement: Why We Need Qualitative Methods for Validation in the Quantitative Era of AI

Thursday, 18th July - 14:45: Cito Special Symposium: The future of educational measurement: Integrating AI and psychometrics (Vencovského aula) - Symposia

Mr. Max van Haastrecht (Leiden University)

Artificial intelligence (AI) is making its presence felt in educational environments. This is not surprising, as machine learning algorithms, like earlier educational technologies, open up new and promising avenues for teaching and learning. However, integrating technology into educational environments places a burden on researchers and practitioners to understand the technology they are using. Validity argumentation, therefore, needs to be adapted to attend to the full socio-technical complexities introduced by innovating educational measurement. Perhaps counterintuitively, we concurrently argue that the increased use of AI means we need to focus more, not less, on the use of qualitative methods when validating educational measurement. Since educational technologies tend to abstract away student behaviour, teachers increasingly have to rely on assessing student data, rather than observing students directly. We need to counter this abstraction with methods that deepen our understanding of educational environments: qualitative methods. Without a thorough understanding of the educational contexts that we are exposing to AI, we risk losing the trust of our students and teachers. We conclude that AI poses a scala of challenges to traditional validity arguments, but that there are clear paths towards holistic validation strategies, even in this challenging era of AI.

Harmonizing ADHD scores by using different methods (Linear equating, Kernel Equating, IRT, Machine Learning)

Thursday, 18th July - 14:45: Cito Special Symposium: The future of educational measurement: Integrating AI and psychometrics (Vencovského aula) - Symposia

Miljan Jovic (University of Twente), Dr. Maryam Amir-Haeri (University of Twente), Dr. Andrew Whitehouse (Telethon Kids Institute), Dr. Stéphanie van den Berg (University of Twente)

A problem that applied researchers and practitioners often face is the fact that different institutions within research consortia use different scales to evaluate the same construct which makes comparison of the results and pooling challenging. In order to meaningfully pool and compare the scores, the scales should be harmonized. The aim of this paper is to use different test equating methods to harmonize the ADHD scores from Child Behavior Checklist (CBCL) and Strengths and Difficulties Questionnaire (SDQ) and to see which method leads to the result. Sample consists of 1551 parent reports of children aged 10-11.5 years from Raine study on both CBCL and SDQ (common persons design). We used linear equating, kernel equating, Item Response Theory (IRT), and the following machine learning methods: regression (linear and ordinal), random forest (regression and classification) and Support Vector Machine (regression and classification). Efficacy of the methods is operationalized in terms of the root-mean-square error (RMSE) of differences between predicted and observed scores in cross-validation. Results showed that with single group design, it is the best to use the methods that use item level information and that treat the outcome as interval measurement level (regression approach).

Using natural language processing to score [short] answers to open-ended items

Thursday, 18th July - 14:45: Cito Special Symposium: The future of educational measurement: Integrating AI and psychometrics (Vencovského aula) - Symposia

Dr. Joost Kruis (Cito Institute for Educational Measurement), Eva de Schipper (Cito Institute for Educational Measurement), Dr. Remco Feskens (Cito Institute for Educational Measurement)

A general trend in testing has been to use questions with a multiple-choice format, and there are many reasons why we often prefer using them over open-ended questions. They are easy to administer and can be automatically scored, and therefore there is no effect from subjectivity during scoring. However, at the same time, they also introduce guessing behaviour and other unwanted response strategies that hinder assessing the true ability of a test-taker. As such, we often see that as the stakes of a test increase, the number of items with an open-ended format in a test increases. The recent surge in (Generative) Artificial Intelligence (AI) has profoundly influenced educational assessment and learning, advertising a seemingly unlimited number of innovative possibilities. One of the areas in which AI shows a lot of promise is the automated scoring of open-ended item responses using natural language processing models, with the aim of providing an automated, more efficient method of assessing student responses. In this project, we investigate pre-processing, feature extraction, and similarity scoring, and how these techniques can be used to score responses to open-ended items with optimal accuracy and consistency. We evaluate our model on the responses to open-ended items for the digitally administered Dutch Central Exams. Other aspects of this project discussed will include the challenges of analysing inconsistent, noisy student data and, perhaps more importantly, the essential ethical considerations that should be considered when implementing this technique in automated systems.

Psychometrics as the origin of statistics as the origin of data science

Thursday, 18th July - 14:45: Cito Special Symposium: The future of educational measurement: Integrating AI and psychometrics (Vencovského aula) - Symposia

Dr. Ivailo Partchev (Cito Institute for Educational Measurement)

Time and technological progress change all aspects of life. Interestingly, every major invention, from robots over OCR to the internet, seems to bring about a repackaging and rebranding of statistical techniques. But there is Gelman saying that 'whatever you do, somebody in psychometrics already did it long before'. This became particularly striking as data science was confronted with two-way data, the natural habitat of psychometrics. Stepping into the traces of Novick and Gelman, we examine briefly the origin of recommender systems, commercially the most important branch of data science as of today.

Beyond confirmatory: dimensionality assessment in forced-choice data

Thursday, 18th July - 14:45: Dimensionality and IRT (RB 101) - Oral

Mr. Diego Graña (Universidad Autónoma de Madrid), Dr. Rodrigo Schames Kreitchmann (National University of Distance Education), Dr. Miguel A. Sorrel (Universidad Autónoma de Madrid), Luis Eduardo Garrido (Pontificia Universidad Católica Madre y Maestra), Dr. Francisco J. Abad (Universidad Autónoma de Madrid)

Determining the dimensions in item response data holds critical importance in psychological evaluation, yet remains unexplored for Forced-Choice (FC) data. FC questionnaires have been increasingly popular due to their ability to account for relevant response style biases like social desirability. However, the widespread use of a confirmatory framework to model FC data relies on assumptions that might not always be valid, such as the absence of cross-loadings and the correct structure specification. This study aims to determine if dimensionality assessment methods such as the Kaiser rule, empirical Kaiser, Parallel Analysis, and Exploratory Graph Analysis can detect the dimensions of FC data, as a first step towards an exploratory framework. Utilizing Thurstonian IRT, we conducted a Monte Carlo simulation, manipulating as factors: the number of dimensions, variables and response options (generating traditional and graded forced choice), the loadings mean and range, the correlations among dimensions, the inclusion of unequally keyed and unidimensional blocks, and the sample size. We employed the hit rate, bias, and mean absolute error as dependent variables to evaluate the effectiveness of the dimensionality assessment methods in identifying the dimensions. Parallel Analysis was the only method with good performance, both marginally and overall, but the performance also depended on the FC questionnaire design characteristics. This led to recommendations for FC questionnaire construction, such as including heteropolar or unidimensional blocks and maximizing the differences of main loadings within blocks. It also suggests that Parallel Analysis can successfully be applied to FC as a potential source of structural validity.

Semi-parametric Item Factor Analysis

Thursday, 18th July - 15:00: Dimensionality and IRT (RB 101) - Oral

Dr. Camilo Cardenas (The London School of Economics and Political Science), Dr. Yunxiao Chen (The London School of Economics and Political Science), Prof. Irini Moustaki (The London School of Economics and Political Science)

Traditional item factor models offer limited flexibility in modeling item response functions (IRFs), leading to the adoption of semi-parametric alternatives. However, these models are typically constrained to unidimensional latent variables. In this presentation, we propose a novel framework for multidimensional semi-parametric item factor models. Here, the probability of a correct response is expressed as a weighted sum of multiple non-linear, monotonic functions of latent variables, utilizing I-splines (Ramsay, 1988, Stat. Sci). The dependence structure between latent variables remains parametric and is modeled using a Gaussian copula. This extends the unidimensional model proposed by Ramsay & Winsberg (Psychometrika, 1991) to encompass multiple latent variables. Notably, our framework addresses rotational indeterminacy inherent in latent variable models. We estimate model parameters using an iterative stochastic approximation algorithm employing Langevin dynamics to efficiently sample from the (potentially high-dimensional) latent variable space. We showcase the potential of this framework through simulations and real-world applications.

Computation of Model Scores for Multidimensional Item Response Theory Models fitted with the WLS Estimator

Thursday, 18th July - 15:15: Dimensionality and IRT (RB 101) - Oral

Franz Classe (Deutsches Jugendinstitut e.V.), Prof. Christoph Kern (Ludwig-Maximilians-Universität München), Dr. Rudolf Debelak (University of Zurich)

In this paper, we present the R function *estfunWLS* designed for computing model scores for multidimensional item response theory (MIRT) models, particularly multidimensional Graded Response Models, estimated with the Weighted Least Squares (WLS) estimator. The WLS estimator allows fast estimation of intricate MIRT model parameters through the limited information approach. The R function makes it possible to compute model scores, i.e., the first-order derivatives of the objective function, for models fitted with the WLS estimator. This way, the package facilitates rapid execution of numerous parameter instability tests for MIRT models. The efficient computation of parameter instability tests is crucial for various applications, such as model-based recursive partitioning algorithms. Such algorithms may be used to detect groups of subjects exhibiting Differential Item Functioning (DIF) which are not pre-specified but result from combinations of covariates.

We performed a comparative analysis of the performance of parameter stability tests for models fitted with a limited information approach (here: the WLS estimator) using the *lavaan* package vs. those fitted with a full information approach using the *mirt* package. The new approach has a good Type I error rate, high power, and is computationally faster than analysis via *mirt*.

A Bifactor Thurstonian IRT Model for Multidimensional Forced-choice Questionnaires with Negative Wording Effect

Thursday, 18th July - 15:30: Dimensionality and IRT (RB 101) - Oral

Dr. Chia-Wen Chen (The Psychometrics Centre, University of Cambridge), Dr. Luning Sun (The Psychometrics Centre, University of Cambridge), Dr. Fang Luo (Faculty of Psychology, Beijing Normal University)

Multidimensional forced-choice questionnaires are widely used to reduce response biases in personality and value tests. Brown & Maydeu-Olivares (2011) developed a Thurstoinian IRT model to address the ipsativity issue from classical scoring of forced-choice items. As a recommendation for the identification of the Thrstonian IRT model, negatively keyed items, which are usually negatively worded, are commonly employed in test development. Existing literature has demonstrated that negatively worded items contain strong confounding effects and tend to measure a factor that is distinct from the traits they are intended to measure. In this study, we propose a bifactor model in the framework of the Thurstonian IRT model to capture the factor resulting from the negative wording and examine its effects on the performance of the Thurstonian IRT model. A simulation study based on multidimensional forced-choice item pairs revealed that when the negative wording effect was ignored, the recovery of factor loadings and inter-trait correlations was undermined, whereas the estimation of thresholds remained largely intact. As the effect became more salient, greater biases were observed. Notably, the inter-trait correlations appeared to be over-estimated. The model fit was also negatively impacted when the test length was relatively short. A subsequent empirical study in a high-stakes assessment showed that compared to the standard Thurstoinian IRT model, our bifactor model fitted real data better and produced a correlation matrix between the Big Five personality traits closer to that in previous personality studies.

Inferring Mastery of Math Skills Over Time using Power Priors

Thursday, 18th July - 14:45: Priors in Bayesian Models (NB A) - Oral

Dr. Josue Rodriguez (McGraw Hill Education), Dr. Chelsea Krantsevich (McGraw Hill Education), Dr. Gabriela Stegmann (McGraw Hill Education), Dr. Eric Ho (McGraw Hill Education), Dr. Yuning Xu (McGraw Hill Education), Dr. Josine Verhagen (McGraw Hill Education), Dr. Angelica Gonzalez (McGraw Hill Education)

We discuss a Bayesian approach to infer a learner's mastery of skills from longitudinal data using power priors (Ibrahim & Chen, 2000). Online educational assessment systems can generate an abundance of longitudinal learner response data. An important goal is to use these data to infer changes to a learners' knowledge over time. That is, assuming a learner's knowledge is changing over time, to what extent should past data influence inferences about a learner's knowledge today? A student could have previously demonstrated non-mastery but since then "learned" the skill or they previously demonstrated mastery of a skill but are currently susceptible to "forgetting" (e.g., Averell & Heathcote, 2011). Using (anonymized) real-world data from online educational assessments of math skills, we demonstrate (1) how power priors can be used to borrow information from multiple historical datasets in cognitive diagnostic models and (2) the roles of "learning" and "forgetting" in determining the appropriate amount of information to borrow. Specifically, we apply the power prior within a Bayesian framework for the deterministic input, noisy "and" gate model (DINA, de la Torre, 2008) to infer a learner's mastery of math skills over time. We conclude by discussing practical implications of this approach in a classroom setting.

On the prior effective sample size of normal and Dirichlet priors for threshold parameters

Thursday, 18th July - 15:00: Priors in Bayesian Models (NB A) - Oral

Dr. Noah Padgett (Harvard University)

Threshold parameters, or item locations, are a key feature of items when constructing measures. When item factor models are estimated with a Bayesian approach, the prior choice for thresholds can majorly impact the posterior distribution for thresholds, especially at lower sample sizes (e.g., < 500 cases) when the response distribution is skewed. An important consideration when specifying priors is how informative the chosen is for the given model. Descriptions of the informativeness of a prior are often vague, using terms such as “moderately informative.” This common language can lead to a misrepresentation of how informative a prior is within a given context. In this work, I will describe how the informativeness of a prior can be characterized in terms of sample size such that the informativeness can be interpreted as the number of additional cases added to the estimation. The prior effective sample size is approximated via simulation by estimating a Bayesian item factor model with a suitably diffuse ε-information before simulated data. The prior effective sample size is then needed to minimize the distance between a target prior and posterior of the model using the ε-information prior (Morita and colleagues, 2008, 2010, 2012). To illustrate, my presentation will compare the prior effective sample size of normal and Dirichlet priors for threshold parameters in item factor analysis. Interpreting priors in terms of sample size will aid the specification of informative priors and provide a common scale by which to evaluate the choice of priors that is more widely accessible.

Bayesian power priors in latent variable models with small samples

Thursday, 18th July - 15:15: Priors in Bayesian Models (NB A) - Oral

Dr. Lihan Chen (McGill University), Dr. Milica Miočević (McGill University), Dr. Carl Falk (McGill University)

Researchers are often faced with small sample sizes in their studies due to various practical constraints, such as limited subject pools and prohibitively costly measures. One approach to handling this small sample problem is to perform Bayesian data analysis using informed priors constructed from historical data, which can improve the efficiency, power, and convergence rate of the analysis. However, when historical data are not fully exchangeable with the current sample, such informed priors can bias the results. To allow for the partial borrowing of historical data such that non-exchange data do not overwhelm the current sample, a *power prior* approach (Ibrahim & Chen, 2000) can be used to downweigh the historical dataset. Recently, Golchi (under review) proposed using normalized Mahalanobis distance to assign an individual weight to each past observation, giving lower weights to observations that differ more greatly from the current sample, rather than uniformly downweighing the entire dataset. This objective weight approach yielded promising results in mediation analysis with manifest variables, showing an overall advantage compared to the noninformative diffuse prior (Miočević & Golchi, 2022). In the current study, we further investigate the objective weight power prior approach in a simulation study, extending its application to two latent variable models–a latent growth curve model and a latent mediation model–to compare its performance with diffuse priors and frequentist maximum likelihood estimation. Our results demonstrate some advantages of power priors, but also bring up some cautionary notes about the limitations and potential improper application of this approach.

Zero-And-One-Inflated Generalized Dirichlet Distribution to Estimate The Order of Resource Allocation

Thursday, 18th July - 15:30: Priors in Bayesian Models (NB A) - Oral

Mr. Taisei Wakai (The University of Tokyo), Prof. Yasunori Kinosada (Shizuoka Institute of Science and Technology), Prof. Yoshiya Furukawa (Fukuoka University), Prof. Ken'ichiro Nakashima (Hiroshima University)

Experiments in social psychology often use tasks that involve the allocation of resources (e.g., money or time), such as dictator games or ultimatum games. Typically, the compositional data from these tasks is analyzed based on the amount of allocated resource. However, the order of resource allocation has rarely been studied, despite its importance in understanding human behavior. This may be due to the lack of means for inferring the order of allocation. This study proposes an extended version of generalized dirichlet distribution (Conner & Mosimann, 1969) to estimate the allocation order trend using compositional data obtained from experiments. The accuracy of the estimation is verified through simulation, and the usefulness of the proposed model is demonstrated through real data analysis. In addition, we will discuss the possibility of improving the measurement validity of constructs by developing novel statistical models that analyze data from a different perspective.

Prognostic score methods for the estimation of average causal effects

Thursday, 18th July - 14:45: Topics in Causal Inference 2 (NB B) - Oral

Mrs. Chamika Porage (Uppsala University), Prof. Ingeborg Waernbaum (Uppsala University)

The Prognostic Score (PGS) is a function of observed covariates that summarizes covariates for response. In our research, we introduce the concept of a Full Prognostic Score (FPGS), a vector comprised of individual prognostic scores. Under effect modification, we prove FPGS is sufficient for confounding adjustment, and implemented FPGS is sufficient for the estimation of the average causal effect. In this context, we initially obtain the estimated PGS by fitting both linear and non-linear regression techniques. When determining the average treatment effect, we apply FPGS in semi-parametric estimators including regression imputation and Targeted Maximum Likelihood Estimation (TMLE). The finite sample properties of estimators are compared through three simulation studies. In an empirical study, we analyze data from the National Health and Nutrition Examination Survey (NHANES,2007-2008) to assess the effect of smoking on blood lead levels. Based on the findings of FPGS estimators, the mean squared error associated with the linear regression imputation estimator and TMLE estimator, which incorporates linearly regressed PGS, is smaller than that of other estimators.

Monotonicity matters: Evaluating heterogeneous treatment effect estimation via nonparametric regression tree methods

Thursday, 18th July - 15:00: Topics in Causal Inference 2 (NB B) - Oral

Mr. Graham Buhrman (University of Wisconsin-Madison), Xiangyi Liao (University of Wisconsin - Madison), Prof. Jee-Seon Kim (University of Wisconsin-Madison)

Interest in estimating heterogeneous treatment effects has substantially increased in recent years. Treatment heterogeneity describes the case when individuals are differentially affected by an intervention or exposure according to their characteristics, and accurate estimation of these differential effects can support plans for optimal resource allocation and/or personalized interventions. However, the manner by which a treatment effect varies across individuals' characteristics can take many forms and is typically unknown to researchers. For instance, the effect of math tutoring on students' test scores might vary across students' prior math scores as a negative parabola, meaning that students who benefit most do not have particularly high or low prior scores. Such "Goldilocks" effects and other complex treatment functions have motivated the use of nonparametric regression techniques which make few or no assumptions about the true data-generating model. Specifically, nonparametric regression tree methods such as Bayesian additive regression trees and random forests have been shown to accurately recover heterogeneous treatment effects. While previous studies have compared the performance of different nonparametric methods in varying data contexts and across different true data-generating models, few studies have explicitly explored how the heterogeneous treatment function's monotonicity (or lack thereof) impacts the performance of nonparametric regression tree methods. This study aims to 1) detail how monotonicity affects the accuracy of nonparametric regression tree estimates of treatment heterogeneity and 2) provide practical recommendations and guidance for the implementation of nonparametric regression tree methods when there is suspected non-monotonicity in the treatment effect function.

Disentangling Person-Dependent and Item-Dependent Causal Effects: Applications of Item Response Theory to the Estimation of Treatment Effect Heterogeneity

Thursday, 18th July - 15:15: Topics in Causal Inference 2 (NB B) - Oral

Mr. Joshua Gilbert (Harvard University), Luke Miratrix (Harvard University), Mr. Mridul Joshi (Stanford University), Dr. Ben Domingue (Stanford University)

Analyzing heterogeneous treatment effects (HTE) plays a crucial role in understanding the impacts of educational interventions. A standard practice for HTE analysis is to examine interactions between treatment status and pre-intervention participant characteristics, such as pretest scores, to identify how different groups respond to treatment. This study demonstrates that identical patterns of HTE on test score outcomes can emerge either from variation in treatment effects due to a pre-intervention participant characteristic or from correlations between treatment effects and item easiness parameters. We demonstrate analytically and through simulation that these two scenarios cannot be distinguished if analysis is based on summary scores alone. We then describe a novel approach that identifies the relevant data-generating process by leveraging item-level data. We apply our approach to a randomized trial of a reading intervention in second grade, and show that any apparent HTE by pretest ability is driven by the correlation between treatment effect size and item easiness. Our results highlight the potential of employing measurement principles in causal analysis, beyond their common use in test construction.

Strengthening causal claims from educational studies: A sensitivity analysis approach

Thursday, 18th July - 15:30: Topics in Causal Inference 2 (NB B) - Oral

Ms. Elaine Chiu (University of Wisconsin-Madison)

Contingency tables are widely used in education and psychometrics to describe associations between two categorical variables. Short of a randomized experiment, the associations that are observed in a contingency table are not causal. Sensitivity analysis is a framework to assess the robustness of the estimated associations from observational studies. There does not exist a sensitivity analysis for contingency tables greater than a 2x2 table. We extend the Rosenbaum sensitivity model to assess the sensitivity of a larger class of test statistics that quantify the association between the rows and the columns of an IxJ table, including the chi-square test for independence, the likelihood ratio test, the global odds ratio test, and linear by linear association test. We apply our method to assess the association between three types of pre-kindergarten (pre-k) care and students' overall math achievement, measured on a discrete scale, from the Early Childhood Longitudinal Study-Kindergarten cohort. After controlling for socioeconomic and demographic factors, we find that the association between pre-k programs and math performance is strong, especially for black and Hispanic female students (two-sided p-values: 0.0034 and 0.0086, respectively), and the observed associations are insensitive up to a Rosenbaum's Γ of 2. Or, roughly speaking, the observed associations are no longer statistically significant at the significance level of 0.05 if there is an unmeasured confounder that simultaneously changes the odds of enrolling into a pre-k program and success in math achievement by a factor of 2.4.

Title: Estimating the Causal Impact of Test Changes on Test Scores in Continuous High-Stakes Assessments

Thursday, 18th July - 15:45: Topics in Causal Inference 2 (NB B) - Oral

Dr. Manqian Liao (Duolingo), Dr. J.R. Lockwood (Duolingo)

High-stakes assessments benefit from continual improvements, such as expanding item banks, optimizing item selection algorithms, adding new features to automated scoring algorithms, refining user interfaces, and expanding construct coverage by adding new item types. Such changes require careful analysis to ensure that score comparability is maintained. For continuously-administered assessments, these analyses can be challenging due to fluctuating test-taker populations that may confound true impacts of test changes. Thus there is a need for analytical methods capable of estimating causal effects of test changes in these contexts. In this study, we consider methods for estimating causal effects of test changes for continuously administered high-stakes assessments. Given the ethical limitations of implementing randomized experiments in high-stakes settings, we adopt two quasi-experimental methods: the posttest-only with nonequivalent groups design and difference-in-differences (DID). In both cases, test takers are classified into “pre-launch” or “post-launch” groups based on the test change implementation date. The first design compares scores across pre/post-launch groups, using a weighting method for group equalization, while DID assesses the difference in score differences between “pre-launch/post-launch” and “pre-launch/pre-launch” repeaters. We examine threats to validity of causal inferences that arise from different types of test changes, particularly emphasizing the potential behavioral shifts among test takers prompted by the visibility or announcement of the changes. We present examples for each type of test change and outline strategies to address the identified validity threats, offering insights into maintaining the comparability and validity of test scores amidst test updates.

Assessing clinical avoidance behavior: A testing framework using process data

Thursday, 18th July - 14:45: Psychometric Applications to Health (NB C) - Oral

Mr. Nico Remmert (Freie Universität Berlin), Dr. Jane Gregory (University of Oxford), Ms. Sewon Oh (University of South Carolina), Prof. Svetlana Shinkareva (University of South Carolina), Dr. Silia Vitoratou (King's College London), Prof. Robert Krause (University of Kentucky), Prof. Steffi Pohl (Freie Universität Berlin)

Many assessments of clinical avoidance behavior rely on self-report measures. These are, however, impacted by response styles, social desirability, or limited awareness, and as such threatening the validity of the measure. We make use of process data to tackle this problem. Instead of self-reports, we assess actual avoidance behavior. Our framework leverages reaction times from computer testing on reactions to the exposure of aversive stimuli. Respondents are exposed to aversive stimuli presented for a specified time interval, with the option to stop the stimuli exposure at any time when they become intolerable. The time until a stimulus is stopped is denoted as endurance time with an upper bound implying full stimulus endurance. The manifest endurance times are analyzed by a one-inflated Item Response Theory (IRT) model for bounded continuous data by Molenaar et al. (2022). With the model, it is possible to estimate a person parameter, denoted as endurance ability, along with four item parameters for each item: endurance discrimination, endurance difficulty, endurance inflation, and item dispersion. The advantages of the framework include its ability to offer a direct measurement of avoidance behavior, modeling flexibility, ethical harmlessness in the testing procedure involving aversive stimuli, and a high reliability.

To demonstrate the framework, an empirical example is presented, focusing on a disorder characterized by decreased sound tolerance, commonly known as misophonia. The study showcases how the proposed framework quantifies reactive avoidance behavior in a clinical context, offering a valuable approach for future research and psychometric assessment.

Investigating measurement invariance of the Simplified Beck Depression Inventory using rating scale tree model

Thursday, 18th July - 15:00: Psychometric Applications to Health (NB C) - Oral

Mr. Farshad Effatpanah (TU Dortmund University), Prof. Manfred Schmitt (RPTU Kaiserslautern-Landau), Prof. Olga Kunina-Habenicht (TU Dortmund University)

Measurement invariance is a crucial consideration in psychological and educational measurement which evaluates the psychometric equivalence of a latent trait across different groups or time points. This property is typically assessed by differential item functioning (DIF). Numerous statistical techniques have been proposed to investigate DIF, including Mantel-Haenszel, logistic regression, likelihood ratio test, multiple-group factor analysis, multiple indicator multiple cause, item response theory (IRT)-/Rasch-based analytical methods, and multidimensional IRT. However, almost all of these methods require a priori specification of two or more groups. This study aims to apply a tree-based global model test for polytomous Rasch models, built on model-based recursive partitioning algorithm (Komboz et al., 2018), to the simplified Beck Depression Inventory (BDI-S) to investigate DIF across age and gender. Unlike the conventional methods, the model does not require a priori specification of groups for detecting DIF. The model splits the sample by subjecting the data to iterative nonlinear partitioning and estimate item difficulty for each split. To explore possible breaches of measurement consistency in the BDI-S, the responses of 4521 German respondents (both clinical and nonclinical) were analyzed using the psychotree package in R. After checking the fit of the data to the Rasch model, the rating scale tree model was estimated. The analysis generated 19 non-predefined DIF nodes, with varying patterns of item difficulties. The results also indicated that age and gender affect the manifestation of depression. Overall, the findings suggest that the model could effectively capture the underlying interaction between the covariates and the BDI items.

The Patient Activation Measure read aloud in a clinical setting

Thursday, 18th July - 15:15: Psychometric Applications to Health (NB C) - Oral

Ms. Magdalena Holter (Medical University of Graz), Prof. Alexander Avian (Medical University of Graz), Prof. Andreas Wedrich (Medical University of Graz), Prof. Andrea Berghold (Medical University of Graz)

Purpose

Patients with temporary or prolonged visual impairment may find it difficult to complete a questionnaire independently. Especially for patients with chronic conditions, patient activation, the knowledge, skills, and confidence to manage one's own health, is critical for disease management. It is usually assessed using self-report questionnaires. The aim was to investigate the psychometric properties of the German Patient Activation Measure® (PAM) administered through an interview in an everyday clinical setting.

Methods

Outpatients diagnosed with the chronic disease macular edema participated in this cross-sectional study. Patient activation, health status, self-efficacy, quality of life and general mood were assessed. Questionnaires were read to patients by four different interviewers. Psychometric properties of the PAM® were investigated using item response theory (IRT), Cronbach's α and trait-trait correlations.

Results

The analysis included 554 patients. The median age was 69 years (IQR: 62–76). Confirmatory factor analysis supported unidimensionality of the PAM$^{©}$. After comparing several IRT models, a generalized partial credit model was selected with fit indices of RMSEA: 0.062 (95% CI: 0.052-0.072), SRMSR: 0.064, TLI: 0.905, CFI: 0.921 and good infit and outfit for most items. Empirical reliability and Cronbach's α were 0.75. Differential item functioning with a small effect was found for confidence in following one's own medical treatments regarding the interviewer (R^2=0.028). Patient activation showed associations with other questionnaires as expected.

Conclusion

The results suggest that the read-aloud PAM® is comparable to the self-administered version in terms of psychometric properties. Objective assessment in an interview setting is possible, but requires good interviewer training.

Meta-regression trees: three innovations

Thursday, 18th July - 14:45: More on AI and Machine Learning (NB D) - Oral

Prof. Elise Dusseldorp (Leiden University)

In the framework of meta-analysis, meta-regression is performed to explain heterogeneity in effect sizes and study characteristics are used as predictors. A disadvantage of meta-regression is that interactions are often ignored. In a previous study, we proposed meta-regression trees to detect homogeneous subgroups of studies with regard to their effect sizes. The resulting meta-regression tree represents interaction effects between the characteristics and the leaves of the tree are the homogeneous subgroups. However, due to the algorithmic nature of the method, the search strategy may result in local optima, confidence intervals of the effects in the subgroups are too optimistic, and the test of the moderator effect(s) is too liberal. We suggested three innovations to address these issues: a permutation test, a unique bootstrap approach, and a smooth sigmoid surrogate strategy for splitting. These innovations were recently implemented in a new version of the R-package metacart. We will illustrate them using simulation and an application to a real meta-analytic data set.

Introducing multilevel lasso models into value-added assessment for variable selection

Thursday, 18th July - 15:00: More on AI and Machine Learning (NB D) - Oral

Ms. Qing Zeng (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China), Dr. Ping Chen (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China)

Value-added (VA) assessment is designed to evaluate the teacher and school effectiveness based on students' academic achievement. One of the key issues is the selection of covariates, which has long fascinated researchers. Currently, most VA studies only use student's prior achievement as a covariate, although there are many variables affecting teacher and school effectiveness (Levy et al., 2019). To better address the covariates selection issue, the multilevel lasso (MLL) model is introduced into the VA assessment. MLL aims to sparse the independent variables in high-dimensional model (i.e., clustered data) that are weakly related to the dependent variable (Schelldorfer et al., 2011). In this study, two covariate selection models, the traditional multilevel models (MLMs) and MLL models, were compared via varying simulation conditions and a real data example of the PISA 2012 math test. Specifically, the manipulated factors included sample size, effect size, number of covariates, and patterns of correlation among predictors at student level. The evaluation criteria consisted of power, type I error, and bias. Simulation results showed that MLL had better control of Type I error than MLMs with a slight decrease in power when more covariates were given and sample sizes were large. MLL and MLMs exhibited the same estimation results for VA residuals. The practical application of MLL was also discussed. In sum, the MLL method is a promising alternative variable selection framework for future VA estimation and development.

AI-based AIG for Mathematics within a Learning Engineering Framework

Thursday, 18th July - 15:15: More on AI and Machine Learning (NB D) - Oral

Prof. Ji Hoon Ryoo (Yonsei University), Hyo Jeong Shin (Sogang University), Seewoo Li (CLASS-Analytics), Jong Kyum Kwon (Gyeongsang National University)

In recent years, the landscape of educational assessment has seen a transformative shift with the rise of Generative AI in Automated Item Generation (AIG). Generative AI promises to streamline and improve AIG by creating questions that challenge learners cognitively while demonstrating a nuanced understanding of the subject matter. This accessibility of generative AI has increased significantly, due to advancements in cloud computing, open-source frameworks, and user-friendly interfaces. This also leverages generative models for crafting AIG questions tailored to specific learning objectives. However, it does not mean that all of subjects would be automated with AIG. Item models of AIG have not been completely generated by generative AI, especially within STEM area. This study elaborates an algorithm of AI-based AIG incorporated with human validation process well-fitted to mathematics and examines the efficiency of item model generation via the algorithm. In addition to the algorithm elaboration with human validation and the examination of the efficiency of finalizing item model generation, we also articulate the process by introducing a system, called CLASS-Analytics, integrating LLM capabilities, particularly the Retrieval-Augmented Generation (RAG) model within a learning engineering framework. CLASS-Analytics autonomously validates the question for syntax errors and, upon educator approval, auto-reports the question's difficulty and educational utility within the system interface. This meticulously orchestrates workflow harnesses the potential of advanced generative models and integrates seamlessly with CLASS-Analytics, revolutionizing AIG question creation. Integrating generative AI into AIG aligns with strategic goals aimed at enhancing accessibility, efficiency, and innovation in educational assessment.

An introduction to anytime-valid inference in psychology

Thursday, 18th July - 15:30: More on AI and Machine Learning (NB D) - Oral

Mark Van Lokeren (Imperial College London)

Increasing evidence suggests that a considerable fraction of published research in psychology has been claimed as irreproducible. In order to achieve statistically significant results researchers resort to questionable research practices. An example is p-hacking, which can be defined as any measure a researcher applies to render a previously non-significant p-value significant. Often it entails that some more data are gathered until the p-value becomes significant. This obviously invalidates the p-value, and error guarantees no longer hold.

Anytime-valid inference has been developed as a new method to solve the above issue. It is based on martingale theory and aims at allowing inference at arbitrary stopping times during the data collection process. As such, it remains valid under continuous monitoring, and optional stopping or continuation.

In this talk, we will discuss the principles of anytime-valid inference by introducing the key concepts of confidence sequences and p-processes. They serve as counterparts to the classical fixed-time confidence intervals and p-values. We will have a look at their properties and show how they can be applied in some examples.

Stop using d′ and start using d_a : Evidence from empirical data on how to measure sensitivity in recognition memory

Thursday, 18th July - 14:45: Psychometric Applications to Cognition and Learning (RB 209) - Oral

Prof. Adva Levi (Tel-Aviv Univeristy), Prof. Caren Rotello (University of Massachusetts - Amherst), Prof. Yonatan Goshen-Gottstein (Tel-Aviv Univeristy)

The replication crisis sounded a warning signal that many scientific discoveries may be false. One particular cause for such false discoveries is the use of accuracy measures that yield high rates of Type I errors. Here, we focus on binary tasks, as in yes-no recognition. Rotello et al. (2008) used computer simulations to show that due to the incorrect assumption of equal-variance of lures and target, d´, H-F and other common measures typically yields high rates of Type I errors for iso-sensitive conditions that differ in bias. Importantly, we demonstrate the advantage of using the less-known signal-detection measure, d_a. We previously showed, using computer simulations as well, that d_a yields much lower Type I error rates when computed in unequal lure-and-target variance scenarios. However, in order to fully examine the validity of d_a, empirical data is required. By using implied base-rate manipulation for word-recognition task across two conditions, we were able to measure performance in iso-sensitive conditions and compute d_a (and other common measures). Our results show a substantially lower false discovery rates using d_a, compared to d´, H-F or any other measure tested.

Intersectional Approaches to Understanding Heterogeneity of Cognition Aging: A Data-Driven Exploration

Thursday, 18th July - 15:00: Psychometric Applications to Cognition and Learning (RB 209) - Oral

Dr. Sunmee Kim (University of Manitoba)

The trajectories of cognitive aging among older adults often vary with diverse aspects of identity and social contexts, yet intersectionality literature has paid little attention to how these moderate cognitive functions over time. Moreover, conventional longitudinal methods for analyzing aging or developmental trajectories (e.g., linear mixed models, growth curve models) have limitations in examining the complexities of these moderators, mainly because they require predefined interaction terms for various moderators prior to data analysis. To address these gaps, we utilize generalized linear mixed-model (GLMM) trees, a flexible data-driven method that combines recursive partitioning for identifying heterogeneous subgroups within data and GLMM for the analysis of aging trajectories. Based on a given list of potential moderators, the method automatically identifies the combination of these moderators that explain the heterogeneity in cognitive function changes over time. This is particularly useful in contexts with numerous potential moderators, each of which has multiple subcategories or exists in a continuous format, making the pre-specification of all possible combinations is impractical. By applying this method to the University of Michigan's Health and Retirement Study data on cognitive aging and Canadian Longitudinal Study on Aging, we demonstrate how GLMM Trees can identify previously unrecognized combinations of moderators, thus revealing the underlying complex intersectional dynamics of socioeconomic variables in cognitive aging patterns.

Score development for the NIH Infant and Toddler Toolbox Gaze

Thursday, 18th July - 15:15: Psychometric Applications to Cognition and Learning (RB 209) - Oral

Dr. Lihua Yao (Northwestern University Feinberg School of Medicine)

The NIH Infant and Toddler ("Baby") Toolbox (NBT) is a groundbreaking measurement tool designed for children aged 1 through 42 months. NBT gaze detection uses a front-facing camera built into a standard iPad Pro, which makes it non-invasive, inexpensive, and easy to administer. The resulting data can be used to determine the proportion of looking to one of two sides (left/right preference), identify a "look" to trigger gaze contingent responses, and discriminate between on/off screen looking for habituation designs. Eye gaze positions include X and Y coordinates captured at 60 frames per second (fps), and images of the participants are taken at 20fps. NBT uses gaze to assess performance in language comprehension, approximate number system, mental transformation, learning and memory, and processing speed.

This presentation will discuss the scoring rules developed using different models. First, cluster analysis was conducted using 11910874 captured records with 620 unique test takers. Item level scoring rules were developed from the captured location of the gaze by comparing it with the supposed correct locations of the gaze. 18 item level scores were created for each test taker, and multidimensional graded response models were applied to the data for the purpose of finding the best fitted model for scoring purpose. Finally, machine learning models were applied to predict the gaze location, because there were many instances where the gaze location was not captured.

Detecting test speededness based on Schwarz information criterion

Thursday, 18th July - 14:45: Response Processes and Test Taking Behaviors (RB 210) - Oral

Dr. Jing Lu (Northeast Normal University), Dr. Jiwei Zhang (Northeast Normal University)

Change point analysis (CPA) is utilized to identify the existence of points within a response sequence that can divide the sequence into two segments with distinct statistical properties. Various aberrant behaviors, such as test speededness, cheating behavior, and examinees' performance decline, have been detected using CPA. In this paper, we propose three CPA approaches based on Schwarz information criterion (SIC): response data only, response time (RT) data only, and the combination of response and RT data, to detect the prevalent test speededness in time-limit tests. To comprehensively investigate the efficiency and accuracy of the proposed CPA approaches based on SIC (abbreviated as SIC-CPA), three simulation studies were conducted using different datasets: responses, RTs, and a combination of both. Simulation results demonstrate that our proposed SIC-CPA methods can effectively enhance the power of change point detection and reduce Type I errors. Moreover, the SIC-CPA method combining responses and RTs outperforms the SIC-CPA method based solely on RTs, and the latter is significantly superior to the SIC-CPA method based solely on responses. In addition, compared to the cumbersome likelihood ratio test and Wald test used in Shao et al. (2016) and Cheng and Shao (2022), the proposed methods avoid the complex computation and greatly reduces the time required to identify locations of test speededness. A real data analysis was conducted to show the application of the proposed approaches.

Accounting for rapid guessing in competence estimation in large-scale assessments

Thursday, 18th July - 15:00: Response Processes and Test Taking Behaviors (RB 210) - Oral

Eva Zink (Leibniz Institute for Educational Trajectories, Bamberg), Jana Welling (Leibniz Institute for Educational Trajectories, Bamberg), Timo Gnambs (Leibniz Institute for Educational Trajectories, Bamberg)

Large-scale assessments provide valuable information on achievement gaps between subgroups in the population, informing politics and shaping educational decisions. However, recent research suggests that some gaps could at least partly be attributed to differences in test-taking motivation than in achievement. Existing approaches for mitigating the distorting effects of rapid guessing on competence estimates focus mainly on point estimates of ability, disregarding the fact that competence analyses are often based on plausible value estimates. This study used data from the National Educational Panel to demonstrate the impact of rapid guessing on achievement estimation using plausible values, analyzing the achievement gap between different types of schools. Four models accounting for rapid guessing in the plausible values estimation were compared: (1) a *baseline mode*l did not account for rapid guessing while including the group variable in the background model, (2) a *person level model* incorporated response time effort as an additional covariate in the background model, (3) a *response level model* filtered all responses with item response times lower than a predetermined threshold and a (4) *combined model* merged the person level model and the response level model entailing filtering fast responses while including response time effort as a covariate in the background model. Combining the results of a preliminary simulation study and this study, recommendations for future research practice are given to improve achievement estimation.

Enhancing Tree-Based Models for Detecting Test-Taking Behaviors: Integrating Response Time and Process Variables

Thursday, 18th July - 15:15: Response Processes and Test Taking Behaviors (RB 210) - Oral

Ms. Tamlyn Lahoud (University of Georgia), Dr. Shiyu Wang (University of Georgia), Ms. Constanza Mardones Segovia (University of Georgia), Ms. Eunkyoung Cha (University of Georgia), Prof. Allan Cohen (University of Georgia), Dr. Yasemine Copur-Gencturk (University of Southern California)

Tree based models are widely employed in classification tasks, particularly in detecting test-taking behaviors or response styles. However, conventional approaches often rely solely on participants' responses to questions, potentially overlooking crucial information. This study seeks to enhance traditional tree-based models by integrating response time—a key variable in the response process extensively studied for identifying suspicious test-taking activities. Moreover, we aim to augment the model by incorporating additional factors reflecting the testing process, as well as characteristics of participants and questions. Our proposed approach will undergo rigorous evaluation. Initially, a simulation study will assess the efficacy of our estimation method and classification accuracy across various conditions. Subsequently, we will apply the proposed methodology to real-world data to detect teacher bias in grading student tests. Specifically, we will analyze grading behavior patterns to identify inconsistencies and ascertain whether teachers exhibit rating bias. This analysis will utilize a range of factors, including teachers' rating of the correctness of solutions, students' mathematical ability and effort, as well as scales measuring teachers' beliefs and dispositions. Notably, response times for rating each solution will also be recorded. Additionally, student-related information such as test performance and demographic data will be considered. By integrating response time and other pertinent factors into tree-based models, our research aims to provide a more comprehensive framework for detecting test-taking behaviors and addressing issues of bias in educational assessments.

Bayesian Factor Mixture Modeling with Response Time for Detecting Careless Respondents

Thursday, 18th July - 15:30: Response Processes and Test Taking Behaviors (RB 210) - Oral

Ms. Lijin Zhang (Stanford University), Dr. Esther Ulitzsch (Centre for Educational Measurement (CEMO), University of Oslo), Dr. Ben Domingue (Stanford University)

Careless respondents inject noise into the response data which subsequently compromises estimation accuracy and model fitting in Confirmatory Factor Analysis (CFA). Traditionally, researchers have depended on the use of reverse-worded questions to detect inattention. However, the rise of online data collection platforms has made response time an appealing target for detection of inattentive respondents in CFA. We introduce a Bayesian factor mixture model that utilizes response time to identify careless respondents. By simultaneously modeling responses and response times, this approach effectively identifies individuals who exhibit short response times alongside a low correlation in their item responses. Through simulation studies, we found that: (1) the proposed model achieves high estimation accuracy of key model parameters such as loadings and intercepts; (2) it demonstrates high accuracy and sensitivity in correctly classifying respondents as either careful or careless; and (3) it maintains classification error rates at an acceptable level. Additionally, an empirical study tests applicability of our model in real-world scenarios, comparing its performance against the traditional method based on reverse-worded questions. The results underscore the effectiveness of the proposed model in excluding careless responses and improving model fit, highlighting its practical advantages.

Modeling careless responding in experience sampling data

Thursday, 18th July - 15:45: Response Processes and Test Taking Behaviors (RB 210) - Oral

Milla Pihlajamäki (KU Leuven), Dr. Gudrun Eisele (KU Leuven), Prof. Ginette Lafit (KU Leuven), Prof. Olivia Kirtley (KU Leuven), Prof. Inez Germeys (KU Leuven)

As the experience sampling method (ESM) is becoming an increasingly popular data collection method in various fields including psychology and psychiatry, it is important to identify potential threats to ESM data quality. One such threat is careless responding, i.e., answering items without sufficient regard to the item content. Although this has been studied in cross-sectional studies, important questions about the dynamics of careless responding over time remain unanswered. The goal of the present study is to use Latent Markov models (LMMs) to model careless responding over time. Specifically, we aim to identify states of careless responding based on previously proposed indices of careless responding, and to model how individuals switch between these states using initial and transition probabilities. We also aim to quantify the relationship between the initial/transition probabilities and both occasion-level (context, affect) and person-level (personality, psychopathology) covariates that have been suggested to influence careless responding in previous literature. In a postregistered study, we use three datasets covering different populations (students, community sample, clinical sample) from the open EMOTE database (Kalokerinos et al., in preparation) to assess careless responding in samples that are representative of those in typical ESM studies. By applying a novel statistical approach to study careless responding dynamics, we will provide a nuanced understanding of when careless responding occurs and what occasion- and person-level variables are associated with it. We hope to provide ESM researchers with concrete guidelines on how to consider careless responding in their study design and/or statistical analyses.

Comparing Large-Scale Assessments with the Total Survey Error Approach

Thursday, 18th July - 14:45: Topics in Large-Scale Assessments (RB 211) - Oral

Dr. Peter van Rijn (ETS Global), Dr. Han Hui Por (Educational Testing Service), Daniel McCaffrey (Educational Testing Service), Prof. Indrani Bhaduri (NCERT), Dr. Jonas Bertling (Educational Testing Service)

Large-scale educational survey assessments (LSAs) are vital in measuring student learning outcomes in modern education systems. We present a structured framework for comparing LSAs to facilitate studying the impact of design decisions on the precision of reported results. Key elements of the framework are the sampling design, assessment design, analysis methodology, and reporting. In breaking down the precision of reported results such as mean proficiency for groups, we make use of the total survey error approach (Weisberg, 2005). Here, for example, intraclass correlation, between-school and within-school reliability (Cho et al., 2019), correlations between domains, and background and historical information are considered (Wu, 2010). We illustrate this approach by comparing three LSA's: India's national achievement survey (NAS), the United States' national assessment of educational progress (NAEP), and the programme for international student assessment (PISA). Our framework and comparison results highlight the nuanced ways in which each LSA is tailored to meet the specific needs and challenges of its respective assessment purpose and the populations these assessments aim to serve.

New Standards for Validating AI-Based Educational Assessments

Thursday, 18th July - 15:00: Topics in Large-Scale Assessments (RB 211) - Oral

Dr. Hua-Hua Chang (Purdue University)

The widespread implementation of modern AI technology in educational measurement has sparked concerns regarding its potential adverse effects. Throughout the years, APA and AERA, with significant involvement from the Society, have consistently developed and implemented a series of standards, which remain relevant today. On the other hand, modern technologies have allowed us to measure new constructs that were difficult to measure in the past.

As a former President of the Society (2012-2013), I am eager to initiate a discussion on how the Society can effectively address these challenges. First, we should prioritize educating AI developers about established standards, similar to how self-driving cars follow traffic laws. This ensures reliable, valid, and interpretable AI assessment tools. Furthermore, it is essential to undertake fresh construct validity studies for the novel traits that individuals assert can be quantified through AI.

Certain research inquiries require formulation, including questions such as:

- How can the Society best engage with AI developers to ensure adherence to established standards?
- What specific challenges arise in ensuring construct validity for AI-based assessments?
- How can some established psychometric tools be improved through the integration of AI?
- With the increasing use of AI, how can we ensure that educational assessments remain fair and unbiased for all students?

While technology has enabled AI systems to proficiently conduct physical measurements and quantify tangible attributes in the observable world, challenges persist in the domain of psychological measurements. Addressing these issues requires the essential involvement of psychometricians to make meaningful contributions in this context.

The standard error of equated test scores

Thursday, 18th July - 15:15: Topics in Large-Scale Assessments (RB 211) - Oral

Prof. Wim J. van der Linden (University of Twente)

The standard error of observed-score equating is generally presented as an error in test equating caused by random sampling of the examinees from an assumed population. One of the earliest examples is the standard error for the randomly-equivalent-groups design introduced in Lord (1982). It is argued that the cause of random error in an equating is *not* sampling error but measurement error in the observed scores that are equated. Consequently, to obtain the standard error of an equated score, the only required step is an adjustment of the standard error of the observed score for the impact of the equating, which is shown to be a factor with a simple analytic expression. However, as the standard errors currently in use typically ignore randomness due to measurement error, irrespective of the adjustment, the necessary conclusion is a standard error of equated scores always equal to zero. On the other hand, if we do allow for the presence of measurement error, not only the scores on the old and new form are equated but their standard errors as well.

Examining The Performance of Hybrid And Traditional Multistage Tests Under Different Conditions

Thursday, 18th July - 15:30: Topics in Large-Scale Assessments (RB 211) - Oral

Dr. Cagla ALPAYAR (Sivas Cumhuriyet University), Dr. Celal Deha Doğan (Ankara University), Dr. Duanli Yan (ETS)

This study compares the measurement precision of a hybrid multistage adaptive test combining CAT and pre-assembled modules (H-MST) and traditional computerized multistage adaptive test (MST) designs across various ability distributions and test lengths (24, 36, and 48). Hybrid designs, with the same number of stages, were evaluated against MST as a reference using a 1000-item artificial pool. In H-MST designs, the middle or last stage employed item-level adaptive testing (computerized adaptive test-CAT). Findings indicate that applying CAT at the end of the test, especially in short test lengths, produces more precise results, though this impact diminishes with longer tests. Hybrid designs exhibit an advantage over MSTs, particularly when ability distributions are skewed. The study explored two-, three-, and four-stage designs, concluding that three-stage H-MST designs are more sensitive to the stage where CAT is administered. Conversely, four-stage hybrid designs show similar performances to tests where CAT is applied at different stages, especially in medium and long tests. The effect size of the difference in mean measurement precision decreases with an increase in the number of stages. For more effective evaluation, changes in measurement precision values, calculated based on provisional ability estimations at the end of each stage were analyzed. The most significant decrease in error values occurred in the first two provisional ability estimates, regardless of the design. While the positive effect of the CAT stage on measurement precision values was observed, it diminished with longer tests and more stages.

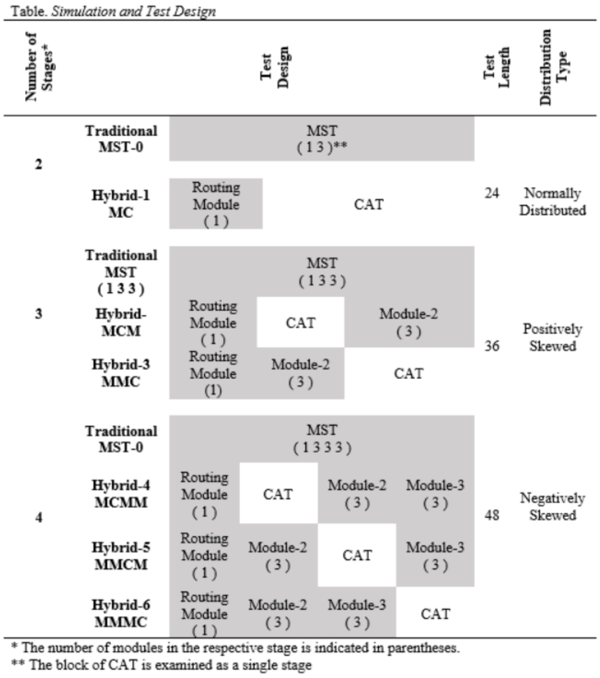


Table. *Simulation and Test Design*

Number of Stages*		Test Design				Test Length	Distribution Type
2	**Traditional MST-0**	MST (1 3)**				24	Normally Distributed
	Hybrid-1 MC	Routing Module (1)	CAT				
3	**Traditional MST (1 3 3)**	MST (1 3 3)				36	Positively Skewed
	Hybrid-MCM	Routing Module (1)	CAT	Module-2 (3)			
	Hybrid-3 MMC	Routing Module (1)	Module-2 (3)	CAT			
4	**Traditional MST-0**	MST (1 3 3 3)				48	Negatively Skewed
	Hybrid-4 MCMM	Routing Module (1)	CAT	Module-2 (3)	Module-3 (3)		
	Hybrid-5 MMCM	Routing Module (1)	Module-2 (3)	CAT	Module-3 (3)		
	Hybrid-6 MMMC	Routing Module (1)	Module-2 (3)	Module-3 (3)	CAT		

* The number of modules in the respective stage is indicated in parentheses.
** The block of CAT is examined as a single stage

Simulation design.png

Importance: A new item parameter that cannot be revealed by single stimulus data

Thursday, 18th July - 14:45: Topics in IRT 2 (RB 212) - Oral

Dr. Safir Yousfi (German Federal Employment Agency), Dr. Susanne Frick (TU Dortmund University)

Many IRT models can be reformulated as latent response (factor) models with intercepts, loadings and latent residual variance as item parameters that determine the latent response at a certain trait level and item thresholds on the latent scale that determine the respective observed response. Consequently, the full latent response formulation of a dichotomous IRT model with 2 parameters (difficulty and discrimination) is overparametrized. Typical identification constraints fix the threshold to zero and dispersion of the latent response or the latent residual to one (DELTA or THETA parametrization in Mplus). The Thurstonian IRT (TIRT) model for ranking data requires only to fix one latent response dispersion in a block. It is shown that latent dispersions can even be compared across blocks by means of a suitable calibration design. This allows to estimate a third item parameter with an interpretation that supplements difficulty and discrimination. It is shown that this new item parameter can be interpreted as importance. An item with higher importance (e.g. “I like the life I live”) can be preferred within a forced choice block although the tendency to agree with this item is lower than the tendency to agree to another item that gets a lower rank (e.g. “I always have a hard time getting up”). The respective substantive psychological item property can only be revealed by comparative data and cannot be identified by single stimulus data. With the knowledge of this new paramter, TIRT FC blocks can be flexibly assembled (for CAT and item banking).

Simulation-based uncertainty estimates for and extension of effective difficulty and discrimination measures for binary item response models

Thursday, 18th July - 15:00: Topics in IRT 2 (RB 212) - Oral

Mr. Peter Johnson (City University of New York), Dr. Jay Verkuilen (City University of New York)

Johnson and Verkuilen (in press) suggested using Fisher information as a measure of effective difficulty and effective discrimination for binary item response theory (IRT) models due to losing the appealing and intuitive meanings of the difficulty and discrimination parameters held in the two-parameter logistic (2PL) model. These metrics show promise for model comparison and parameter interpretability in models that expand, modify, or otherwise differentiate from the 2PL, such as the four-parameter logistic (4PL) family and asymmetric IRT (AsymIRT) models, like the complimentary log-log and logistic positive exponent models. This research aims to solidify the use of these metrics as effect size indices by generating simulation-based uncertainty estimates in order to show the precision of these measures, further promoting the strength of their usage for model comparison and interpretation. This research then looks to extend these measures from binary IRT to use with multidimensional IRT (MIRT) models, aiming to identify how to modify or otherwise adapt these measures for use in a MIRT space, as well as show the precision of the measures there.

Bias-reduced fixed effects estimation of two-parameter logistic models

Thursday, 18th July - 15:15: Topics in IRT 2 (RB 212) - Oral

Prof. Ruggero Bellio (University of Udine), Prof. Nicola Sartori (University of Padua), Prof. Ioannis Kosmidis (University of Warwick)

We illustrate a proper fixed-effects approach for the estimation of the parameters of two-parameter logistic (2PL) models. The idea is to extend to the 2PL setting the bias-reducing estimation approach introduced by Firth (1993) and further studied by Kosmidis & Firth (2009). This amounts to joint estimation of the item and person parameters via the solution of a set of bias-reducing adjusted score equations, which impose an appropriate amount of model-based shrinkage on all the model parameters. Analytic derivations and simulation studies provide evidence that the proposed approach for fixed-effects estimation has satisfactory properties, in many cases outperforming the customarily used approaches based on marginal likelihoods, and quite competitive with alternative approaches that assume a flexible distribution for the person parameters. Some attention will be devoted to the implementation details of the proposal in statistical software, discussing ways to obtain a scalable solution, suitable for fitting the model to data obtained from a large number of subjects.

Should we trust model flexibility? Investigating alternative polytomous IRT models.

Thursday, 18th July - 15:30: Topics in IRT 2 (RB 212) - Oral

Dr. Giovanni Bruno (University of Padua), Prof. Andrea Spoto (University of Padua), Prof. Daniela Di Riso (University of Padua), Prof. Gioia Bottesi (University of Padua)

Measurement models based on Item Response Theory (IRT) have garnered significant recognition in the field of psychological testing. When opting for a polytomous IRT model, several assumptions need to be considered, such as the nature of the measurement scale and the relationship between the latent trait and item responses. A model with fewer assumptions (e.g., the Graded Response Model) provides greater flexibility, making it more attractive for psychological researchers. However, in the investigation of the psychometric constructs, the model selection procedure lacks taking into account the theoretical assumptions of the investigated construct and of the measurement scale. This contribution delves into this topic comparing three IRT models for polytomous items with an increasing level of flexibility (Rating Scale Model, Graded Response Model, and Partial Credit Model) in terms of goodness-of-fit and item parameters, following both a simulative and applied approach using real data. With the present contribution, we underline how the selection of a specific statistical model holds the potential to significantly influence the characterization of a psychological construct, using as example a subclinical population. By addressing the arbitrary nature of model selection in the validation process, this contribution claims that the strength of a "good" IRT model comes not only from the results of statistical testing, but also from its consistency with the foundation theory that led the scale development.

Bit scales for item response theory models

Thursday, 18th July - 15:45: Topics in IRT 2 (RB 212) - Oral

Mr. Joakim Wallmark (Umeå university), Dr. Maria Josefsson (Umeå university), Prof. Marie Wiberg (Umeå university)

Item Response Theory (IRT) is a statistical method used for evaluating test items and assessing test-taker abilities through the analysis of their responses. Since test-taker abilities are latent variables, they are not directly observable in real life. Instead, latent trait scales are constructed when fitting IRT models, but these scales are arbitrary and distances on the scales hold no inherent meaning. In this study, we present a novel concept termed *bit scales*, which are ratio scales grounded in information theory. The latent trait scale from any fitted IRT model can be transformed into a bit scale, enhancing interpretability and facilitating model comparisons. While useful in general IRT contexts, bit scales are exceptionally beneficial when using IRT model fitting algorithms that make minimal or no assumptions about latent trait scale distributions. We demonstrate the usefulness of bit scales through a series of examples using data from the Swedish SAT, and simultaneously introduce a new model for multiple choice data, the *monotone multiple choice model*.

Variable Selection Techniques for Imputation Models

Friday, 19th July - 09:00: ETS Special Symposium: Psychometric innovations in large-scale survey assessments (Vencovského aula) - Oral

Dr. Peter van Rijn (ETS Global)

LSAs calculate group-level statistics, such as means, based on imputations known as plausible values (Mislevy, 1991). Plausible values are also shared with the public for use by secondary analysts in educational research. To minimize congeniality violations in analysis (Meng, 1994), many contextual variables are incorporated into the imputation model, but at the same this can lead to overfitting. Investigating optimal data-reduction procedures for specific conditions is therefore crucial. This study evaluates the use of partial least squares as an alternative to the currently used principal components for variable reduction.

Imputation Models for Special Subpopulations

Friday, 19th July - 09:00: ETS Special Symposium: Psychometric innovations in large-scale survey assessments (Vencovského aula) - Oral

Dr. Usama Ali (Educational Testing Service)

LSAs report subpopulation proficiency distributions, including subpopulations with limited or partial information. Respondents with literacy-related nonresponse, such as those facing language barriers, represent such subpopulations. Excluding these groups from population-level statistics can lead to biased outcomes. Addressing this, the imputation for these subpopulations presents unique challenges that can impact trend reporting across assessment cycles. In this study, we evaluate alternative models for accurately reporting on these special subpopulations.

Linking Error in Large-Scale Assessments

Friday, 19th July - 09:00: ETS Special Symposium: Psychometric innovations in large-scale survey assessments (Vencovského aula) - Oral

Dr. Paul Jewsbury (Educational Testing Service)

- Educational assessments require periodic administration changes, such as transitioning from paper to digital administration. During such transitions, a linking function relating the metrics of the new and previous administration is estimated, but uncertainty in this estimation introduces variance into comparisons between administrations. Similarly, different assessments provided to overlapping populations may be linked, introducing linking or equating error. We introduce new generally applicable variance estimation methods, generalize prior methods to be more widely applicable, and confirm the validity of the methods via simulation. Our methods account for dependencies between linking and other sources of error, complex sampling, and non-linear linking functions, while applying to a wide range of score comparisons and statistics such as means, standard deviations, percentiles and differences.

From traditional to modern methods for the analysis of multi-item measurements

Friday, 19th July - 09:00: Symposium: From traditional to modern methods for the analysis of multi-item measurements (RB 101) - Symposium Overview

Dr. Lubomír Štěpánek (Faculty of Informatics and Statistics, Prague University of Economics and Business; First Faculty of Medicine, Charles University; Institute of Computer Science, Czech Academy of Sciences), Dr. Adéla Hladká (Institute of Computer Science, Czech Academy of Sciences)

The analysis of multi-item measurement plays a crucial role in various fields, including educational assessment, psychological measurement, and health-related outcomes (Martinková & Hladká, 2023). Among others, it comprises the analysis of item functioning, learning the dependence structure between the items and the latent attributes, or text analysis of item wording. Over time, there has been a notable shift from traditional to modern methods in the analysis of multi-item measurements, driven by advancements in technology, statistical methodologies, and a deeper understanding of cognitive processes. This symposium, presented by the members of the Computational Psychometrics (COMPS) Group at ICS CAS and current and former students of Charles University, explores this evolution, highlighting critical aspects of both traditional and modern approaches.

The talk by Lubomír Štěpánek is devoted to some aspects of robust inference behind a very classical concept – the upper-lower index estimating item discrimination. Adéla Hladká will discuss work on innovative estimation algorithms of item functioning accounting for group-specific guessing and inattention. Ján Pavlech will discuss the 4-parameter model in the factor analysis framework and its relationship to the 4-parameter counterpart in the item response theory framework. The paper by Jan Netík focuses on estimating difficulty based on item wording using large language models. Iván Pérez will present an algorithm to learn the structure of a particular type of cognitive diagnostic model called R-RUM and will discuss the results of its identifiability. Throughout the symposium, Dr. Gabriel Wallin will serve as a discussant.

Robust inference for traditional item discrimination index

Friday, 19th July - 09:00: Symposium: From traditional to modern methods for the analysis of multi-item measurements (RB 101) - Symposia

Dr. Lubomír Štěpánek (Faculty of Informatics and Statistics, Prague University of Economics and Business; First Faculty of Medicine, Charles University; Institute of Computer Science, Czech Academy of Sciences)

The upper-lower index measures item discrimination as a difference in the proportions of correct answers between two populations of test-takers. Traditional statistical inference method then assumes identical probabilities of success, i.e., a correct answer to an item, for each test-taker, following binomial distribution. However, in reality, test-takers within a population may have non-identical probabilities of success, leading to a Poisson-binomial distribution rather than a binomial one. Consequently, averaging non-identical probabilities of success across a population and applying traditional methods to compare two populations could bias results. This work addresses the challenges introduced by non-identical probabilities of success and applies the Poisson-binomial distribution to the upper-lower index. By leveraging Popoviciu's inequality, a general formula for a statistic in a robust inference test is derived to compare the proportions of correct answers between the populations. Additionally, using Le Cam's theorem and Chernoff bound, an upper bound for the p-value, i.e., a probability that the difference of counts of successes in both populations is greater than or equal to an observed count, assuming no difference between the populations, is established, which may help to analyze statistical power of the test or minimum required sample sizes, even under the robust scenario assumptions.

The proposed technique is then applied to simulated test data in two populations, providing insights into the statistical properties of this nonparametric approach. The approach considers the realistic scenario of non-identical probabilities of success within populations, offering a more accurate comparison of the proportions of correct answers in testing frameworks.

Innovative iterative algorithms for estimation of item functioning

Friday, 19th July - 09:00: Symposium: From traditional to modern methods for the analysis of multi-item measurements (RB 101) - Symposia

Dr. Adéla Hladká (Institute of Computer Science, Czech Academy of Sciences), Dr. Patrícia Martinková (Institute of Computer Science, Czech Academy of Sciences; Faculty of Education, Charles University)

When the item functioning of multi-item measurement is modeled with three or four-parameter models, parameter estimation may become challenging. Effective algorithms are crucial in such scenarios. This paper explores innovations to parameter estimation in generalized logistic regression models, which may be used in item response modeling to account for guessing/pretending or slipping/dissimulation and for the effect of covariates.

We introduce algorithms for maximum likelihood estimation, including a new implementation of the EM algorithm and a new algorithm based on the parametrized link function. The two novel iterative algorithms are compared to existing methods in a simulation study, which includes nonlinear least squares and a maximum likelihood estimation algorithm that considers constraints on item parameters. Furthermore, we discuss the application of the methods in the context of detecting differential item functioning. Finally, we examine the practical implementation of these methods in the difNLR package (Hladká & Martinková, 2020), accompanied by real-world examples for demonstration.

Three and four-parameter IRT model in factor analysis framework

Friday, 19th July - 09:00: Symposium: From traditional to modern methods for the analysis of multi-item measurements (RB 101) - Symposia

Mr. Ján Pavlech (Institute of Computer Science, Czech Academy of Sciences), Dr. Patrícia Martinková (Czech Academy of Sciences and Charles University)

This work proposes a 4-parameter factor analytic (4P FA) model for multi-item measurements composed of binary items as an extension to the dichotomized single latent variable FA model. We provide an analytical derivation of the relationship between the newly proposed 4P FA model and its counterpart in the item response theory (IRT) framework, the 4P IRT model. A Bayesian estimation method for the proposed 4P FA model is provided to estimate the four item parameters, the respondents' latent scores, and the scores cleaned of the guessing and inattention effects. The newly proposed algorithm is implemented in R and Python, and the relationship between the 4P FA and 4P IRT is empirically demonstrated using real datasets.

Fine-tuning language models to predict item difficulty from wording

Friday, 19th July - 09:00: Symposium: From traditional to modern methods for the analysis of multi-item measurements (RB 101) - Symposia

Jan Netík (Institute of Computer Science, Czech Academy of Sciences; Faculty of Education, Charles University), Filip Martinek (Institute of Computer Science, Czech Academy of Sciences), Dr. Patrícia Martinková (Institute of Computer Science, Czech Academy of Sciences; Faculty of Education, Charles University)

In the domain of educational assessment, crafting items with robust psychometric properties poses significant challenges, especially when pretesting on a pilot population is not feasible. This necessitates reliable methods for estimating difficulty (and possibly other parameters) based solely on item wording. Traditionally, this involves extracting a wide array of theory-driven text features—ranging from basic descriptive statistics to readability indices—as predictors of item difficulty. To derive these text features, item wordings must first undergo extensive preprocessing, which results in a loss of crucial information (e.g., due to lemmatization).

Recently, the focus has shifted towards predictors based on word embeddings, for instance, to better capture the semantics (Štěpánek et al., 2023). However, reflecting the advent of large language models (LLMs) such as transformers, exploring their adaptation for item difficulty prediction presents a promising opportunity. Although these models were originally trained on large corpora of textual data for tasks like masked text prediction, we can leverage the phenomenon of transfer learning and fine-tune these pre-trained LLMs for the task of item difficulty prediction. Thus, we may benefit from the nuanced language representation of modern LLMs without any loss of information along the way and without the need for any separate statistical model.

In this work, we propose and test an innovative approach that utilizes the fine-tuning of pre-trained LLMs to estimate item difficulty from wording. By integrating these modern LLMs, we aim to achieve more accurate predictions of item characteristics, potentially aiding in the process of educational assessment development and evaluation.

On the identifiability and structural learning of R-RUM models.

Friday, 19th July - 09:00: Symposium: From traditional to modern methods for the analysis of multi-item measurements (RB 101) - Symposia

Iván Pérez (Institute of Computer Science, Czech Academy of Sciences; Institute of Information Theory and Automation, Czech Academy of Sciences; Faculty of Mathematics and Physics, Charles University), Dr. Patrícia Martinková (Czech Academy of Sciences), Jiří Vomlel (Institute of Information Theory and Automation, Czech Academy of Sciences)

Cognitive diagnostic models (CDMs) are discrete latent variable models that have been shown to be very useful in educational and psychological measurement for learning the dependence structure between the items and the latent attributes. In this work, we focus on a particular CDM, called the Reduced Reparameterized Unified Model (R-RUM). Our interest in this model is mainly due to the fact that R-RUM is also known in the area of probabilistic graphical models, it corresponds to a particular class of Bayesian networks called BN2A due to its structure. These models have become popular due to their interpretability and flexibility. The identifiability of CDMs is a fundamental prerequisite for valid statistical inference. We will discuss some new results that we have obtained regarding the identifiability of the parameters of R-RUM and present a novel algorithm for learning the structure of these models.

Combining item responses and paired comparisons in small sample contexts

Friday, 19th July - 09:00: Topics in IRT 3 (NB A) - Oral

Dr. Nathan Zoanetti (Australian Council for Educational Research), Dr. Ray Adams (Australian Council for Educational Research), Dr. David Jeffries (Australian Council for Educational Research), Dr. Dan Cloney (Australian Council for Educational Research)

Unavoidably small test taker samples can hinder measurement processes like item calibration and scale equating due to a lack of precision. One potential solution, when it is not possible to collect additional item responses, is to include the items in a paired comparisons study. Following this, the paired comparisons judgements and available item responses and can be combined in a scaling model, potentially increasing the precision of item parameter estimates.

This research demonstrates, through simulations, how data from both item responses and from paired comparisons judgements can be combined and calibrated onto the equivalent of a Rasch (1960) scale using the Bradley-Terry-Luce (1952; 1959) model. Simulation conditions covered a range of sample sizes, paired comparisons exposures per item, and test lengths. Test lengths were varied to enable consideration of the cost of collecting sufficient paired comparisons data. The extent to which the generating item parameters were recovered across simulation conditions was evaluated using mean absolute deviation metrics.
Results show conditions under which the combination of these two data generating mechanisms could be beneficial in some measurement contexts.

Combining item responses and paired comparisons of item difficulty shows some promise when neither data collection is adequate on its own. Several assumptions remain to be tested in practice, including whether the two data generating mechanisms are driven by the same underlying latent trait and whether the average discrimination of the two processes is equivalent. Further research in this area would be valuable for measurement practitioners.

Modeling Process Data with Explanatory Item Response Models

Friday, 19th July - 09:15: Topics in IRT 3 (NB A) - Oral

Dr. Susanne Frick (TU Dortmund University), Miriam Fuechtenhans (University of Kent), Prof. Anna Brown (University of Kent)

Process data are collected along with the responses of primary interest and provide information about the response process. Modelling approaches for process data are often tailored to the specific type of data collection. In this study, we model process data simply with explanatory item response models with appropriate distributions. We illustrate how these models can inform test construction by applying the approach to the case of impression management in high-stakes situations. The empirical objective of this study is thus to investigate how item and person characteristics manifest in response editing in questionnaires.

We conducted a re-analysis of 6 datasets, all of which represent responses to rating scale and forced choice questionnaires and contain a manipulation of stakes, with process data such as response latency and number of clicks. We derived item predictors from desirability ratings obtained from separate samples. To each outcome variable, we fitted explanatory item response models, with log-normal distributions for response latencies and Conway-Maxwell-Poisson distributions for number of clicks.

We found shorter response times for forced-choice blocks of items that were less well matched in desirability, although this was significant in only one study. The effect of ambiguous items differed between rating scale and forced-choice. However, most interactions of the item covariates with stakes were not significant.

From a psychometric perspective, this study can inform further psychometric developments for the analysis of process data. From a practical perspective, the results of this research can inform the development and evaluation of fake-resistant assessments.

A Mixture-IRT Model for Carless Responding with Flexible Assumptions

Friday, 19th July - 09:30: Topics in IRT 3 (NB A) - Oral

Irina Uglanova (IPN – Leibniz Institute for Science and Mathematics Education), Prof. Gabriel Nagy (IPN – Leibniz Institute for Science and Mathematics Education), Dr. Esther Ulitzsch (Centre for Educational Measurement (CEMO), University of Oslo)

Careless and insufficient effort responding (C/IER) occurs when respondents provide their responses without paying attention to the content of the items. Ignoring careless responses potentially leads to bias in parameter estimates and threatens the validity of conclusions drawn from surveys. Mixture-IRT models for C/IER are based on specific assumptions about the structure of C/IER, commonly assuming careless responses to be random. We present an extended mixture-IRT model for C/IER with a careless mixture component that subsumes previous developments as special cases. The proposed model is capable to identify a variety of C/IER behavior, covering random response, respondent-specific category preferences, and straight lining. In an extensive simulation study, we evaluate questionnaire characteristics required for trustworthy C/IER identification (scale length, reliability, item heterogeneity, and the presence of reversed items) and showcase the model's ability to handle various C/IER behaviors. From our results, we derive guidelines for model application and questionnaire design facilitating C/IER identification. In an empirical application, we compared the model's identification of C/IER across four online data collection platforms, administering a Big 5 inventory. The results support the validity of the model's conclusions on C/IER in two aspects. First, the model-implied C/IER rates across data collection platforms for all Big 5 traits were consistent with previous research on the platforms' data quality. Second, posterior C/IER class probabilities showed agreement with multiple behavioral C/IER indicators.

The Item Response Warehouse (IRW)

Friday, 19th July - 09:45: Topics in IRT 3 (NB A) - Oral

Dr. Ben Domingue (Stanford University), Dr. Klint Kanopka (New York University), Mika Braginsky (Stanford University), Ms. Lijin Zhang (Stanford University), Lucy Caffrey-Maffei (Stanford University), Radhika Kapoor (Stanford University), Yiqing Liu (Stanford University), Prof. Susu Zhang (University of Illinois, Urbana-Champaign), Mike Frank (Stanford University)

In contrast with some other quantitative disciplines, psychometrics has a relative paucity of data. The IRW (Item Response Warehouse) is designed to change that through the collection and standardization of a large volume of item response datasets. We describe key elements of the data standardization process and then offer a brief description of the over 200 datasets already in this early iteration of the IRW. We describe the resources available for accessing the data including both a website (https://datapages.github.io/irw/) and API-based access and offer example code illustrating how to download data from the IRW and use it in standard psychometric analyses. We then document next steps that we anticipate taking with the IRW and describe ways that it could be utilized in future research projects.

Standardized person-fit statistics for tests with polytomous items

Friday, 19th July - 10:00: Topics in IRT 3 (NB A) - Oral

Prof. Kylie Gorney (Michigan State University)

Recent years have seen a growing interest in the development of person-fit statistics for tests with polytomous items. Some of the most popular person-fit statistics for such tests belong to the class of standardized person-fit statistics, T, that is assumed to have a standard normal null distribution. However, this distribution only holds when (a) the true ability parameter is known and (b) an infinite number of items are available. In practice, both conditions are violated, and the quality of person-fit results is expected to deteriorate. In this paper, we propose a new correction for T that simultaneously accounts for the use of an estimated ability parameter and the use of a finite number of items. Our simulation study reveals that the new correction tends to outperform not only the original statistic T, but also an existing correction for T that was proposed by Sinharay (2016). Therefore, the new correction appears to be a promising tool for assessing person fit in tests with polytomous items.

Adaptively regulated interim theta bounds for optimizing CAT item selection

Friday, 19th July - 09:00: Topics in Adaptive Testing (NB B) - Oral

Dr. Sung-Hyuck Lee (Graduate Management Admission Council), Dr. Kyung (Chris) Han (Graduate Management Admission Council)

In computerized adaptive testing (CAT), selecting the most relevant item based on specific criteria—such as the maximized information function or the item's difficulty level—is a crucial process. The interim theta estimate, which represents the most current latent trait score derived from previously observed responses, is the primary factor used to evaluate these criteria. In the early phases of a CAT, however, these interim theta estimates can exhibit significant variability, which may result in suboptimal or, in some instances, unfeasible item selections. To address these challenges, developers of CAT-based testing programs often implement a variety of tweaks to their CAT algorithms. These modifications can range from restricting the bounds of the theta scale and limiting the maximum allowable changes in interim theta between items to incorporating the use of a Bayesian estimator, either in full or in part. Although these tweaks are known to be useful in improving the feasibility and stability of CAT operations, the effectiveness of these strategies, especially in terms of optimality of CAT, has rarely been examined in a systematic manner. Moreover, these conventional methods often fail to account for the evolving nature of CAT progression. To bridge this gap, this study introduces two new methods for adaptively and progressively regulating the bounds for interim thetas as the CAT progresses: (1) in a direction that diverges and (2) in a direction that converges. Through comprehensive simulation studies, we assess these methods' impacts on CAT optimality, comparing them with established CAT configurations prevalent in practice.

Non-parametric Item Response Theory for Computerized Adaptive Tests

Friday, 19th July - 09:15: Topics in Adaptive Testing (NB B) - Oral

Laura Aspirot (Universidad de la República), Mario Luzardo (Universidad de la República), Leonardo Moreno (Universidad de la República)

In the domain of Item Response Theory (IRT), Computerized Adaptive Tests (CAT) have grown in importance for large-scale educational evaluation, as they have several advantages over traditional tests. Benefits include higher accuracy in feature estimation, shorter application time, and flexibility. Despite the progress and current applications around CAT, many of them are developed using parametric models and estimating Fisher information or Kullback-Leibler distance in a parametric way. No large-scale applications are using non-parametric models and there are few theoretical developments at the non-parametric level. In this work, non-parametric CAT is studied and developed in the case of univariate and multivariate items, and compared with CAT using parametric IRT models. The models addressed, and compared with the parametric model, are the Ramsay model (Ramsay, 1991) and the univariate and multivariate non-parametric isotonic models (Luzardo & Rodríguez, 2015) . The last models are based on the estimation of the inverse of the item characteristic curves (ICC) by a two-stage process. Item selection was implemented for different rules: maximum non-parametric pseudo-information, non-parametric Kullback Leibler, isotonic non-parametric Kullback Leibler and maximum non-parametric pseudo-information truncated. The performance analysis for the estimators and algorithms for parametric and non-parametric models considers the accuracy of the estimate, bias, number of items needed for a predetermined error and item exposure. For simulated data following the non-parametric model, the isotonic model showed that maximum non-parametric pseudo-information truncated overperformed the other options.

Computerized adaptive testing with continuous items: Improving the efficiency of noncognitive assessments

Friday, 19th July - 09:30: Topics in Adaptive Testing (NB B) - Oral

Mr. Wei-Hung Yang (Department of Educational Psychology and Counseling, National Taiwan Normal University), Prof. Yao-Ting Sung (Department of Educational Psychology and Counseling, National Taiwan Normal University), Dr. Yeh-Tai Chou (Research center for educational and psychological testing, national Taiwan normal university)

Computerized adaptive testing (CAT) has been applied to noncognitive assessments using Likert-type items to reduce test length and testing time without sacrificing measurement accuracy. Recently, the use of Visual Analogue Scales (VASs; Hayes & Patterson, 1921) to assess noncognitive constructs (e.g., career interest and work values) has gained popularity due to enhanced reliability (Cook et al., 2001; Krieg, 1999) and responsiveness (Pfennings et al., 1995) of assessments. However, the efficacy of CAT techniques in improving the efficiency of noncognitive assessments with VAS-type items remains unclear. This study applies the continuous rating scale model (Muller, 1987) to VAS response data and evaluates the performance of CAT with VAS-type items in terms of parameter recovery and testing efficiency through simulation studies. The results show that VAS-type CAT provides reasonable standard errors of measurement and significantly reduces test length, thereby enhancing assessment efficiency for measuring noncognitive constructs. The study discusses the implications of these findings and explores potential future applications in psychometrics.

Addressing calibration error in multidimensional adaptive testing: A Bayesian approach

Friday, 19th July - 09:45: Topics in Adaptive Testing (NB B) - Oral

Dr. Aron Fink (Goethe University Frankfurt)

Computerized adaptive testing (CAT) traditionally relies on item parameters derived from calibration studies and treats them as fixed values during the operational CAT phase. However, this ignores the fact that item parameters are only estimates that contain some degree of calibration error, which can lead to underestimated standard errors and potentially biased ability estimates. This is particularly pronounced in multidimensional CAT (MCAT), where multiple abilities are assessed simultaneously. To address this issue, we propose a novel Bayesian algorithm for MCAT that explicitly models item parameter uncertainty during ability estimation and item selection. In a comprehensive simulation study, the performance of the novel method is compared with traditional MCAT approaches. A multidimensional adaptive test that aims to measure three dimensions is simulated. The multidimensional 2PL model is used as the measurement model with a mixture of between-item and within-item multidimensionality. The simulation study is based on a factorial design with the factors calibration sample size (250, 500, 1000), method of accounting for uncertainty in item parameter estimates (none, Bayes), and test length (30, 60, 90). The evaluation criteria are bias and MSE of the three-dimensional ability estimates, each conditional on the true ability levels of the test takers. While the simulation is ongoing, preliminary results suggest that the fully Bayesian approach leads to more accurate ability estimates and reliable standard errors compared to traditional MCAT methods.

Evaluating Models for Item Cloning Variation under Multistage Testing

Friday, 19th July - 10:00: Topics in Adaptive Testing (NB B) - Oral

Dr. Won-Chan Lee (University of Iowa), Dr. Stella Kim (University of North Carolina at Charlotte)

Many of the current large-testing programs, such as Law School Admission Test (LSAT), Medical College Admission Test (MCAT), and Graduate Record Examinations (GRE) have implemented MST as a mode of test administration. These programs yield scores that are used to make high-stake decisions, including college admission, professional certifications and licensing, and educational placements. As a result, ensuring the quality of assessments and accurate estimation of examinees' ability is of utmost importance.

Furthermore, the emergence of AIG, particularly in the context of the recent ChatGPT release, has spurred intensified discussions, and some testing production companies has already begun to create software to build exams using AIG (e.g., Scorpion). However, the current literature offers limited insights into the characteristics of items generated through AIG, particularly in the realm of MST. This study aims to investigate various competing (advanced) IRT models, tailored to capture the nuances among item clones, and assess their performance under MST. The findings of this study will provide valuable insights for practitioners considering the adoption of AIG and the selection of underlying psychometric models for multi-stage testing (MST). The following research objectives will guide the project: 1) Examine potential IRT models that can be used for items generated by AIG in the context of MST; 2) Evaluate the accuracy of the competing models in recovering examinee's ability through a simulation study.

Continuous-Time modeling of bivariate developmental trajectories in accelerated longitudinal designs

Friday, 19th July - 09:00: Longitudinal Designs and Methods (NB C) - Oral

Ms. Nuria Real-Brioso (Universidad Autónoma de Madrid), Dr. Pablo F. Cáncer (Universidad Pontificia Comillas), Dr. Eduardo Estrada (Universidad Autónoma de Madrid)

Recent research has investigated discrete- and continuous-time (CT) approaches to recover univariate developmental processes in Accelerated Longitudinal Designs (ALDs), which allow capturing extended developmental processes with a smaller number of assessments in a shorter time framework. However, researchers are often interested in the temporal development of two intercorrelated processes such as, for example, cognitive and cortical development. In these scenarios, the use of bivariate models is required. In this work, we conducted a Monte Carlo simulation study to assess the performance of CT-Bivariate Latent Change Score models in the context of ALDs with different sampling conditions. We discuss our findings and provide recommendations for effectively applying these models with data obtained through ALDs. We also provide resources to assist researchers in the implementation of these models, including computer code and an empirical example in cognitive development, which demonstrates the utility of CT-BLSC models in analyzing the joint development of cognitive processes during childhood and adolescence.

More On Cross-Lagged Effects – (Mis)Specification in Dynamic Systems Models

Friday, 19th July - 09:15: Longitudinal Designs and Methods (NB C) - Oral

Dr. Charles Driver (University of Zurich)

Model misspecification significantly impacts the validity of inferences drawn from longitudinal analysis, but is often overlooked or addressed superficially. In this presentation, I delve into key issues surrounding model fit testing, detection methods for misspecification (e.g., row-wise gradient contributions, prior-predictive comparisons of auto and cross-correlation functions), and the robustness of inferences.

I illustrate these concepts through a detailed example (as in Driver, 2024), reanalyzing mood data collected from Dutch undergraduate students (Fried, Papanikolaou, & Epskamp, 2022) using continuous-time dynamic modeling. By exploring various model specifications, including considerations such as measurement error, system order, non-linear measurement models, and time representation (continuous vs. discrete), I highlight the sensitivity of inferences to these differences.

Furthermore, I examine detection and adjustment methods for addressing model misspecification, showcasing the versatility of the ctsem software in accommodating diverse specifications and facilitating sensitivity checks across various model configurations. Through this analysis, I underscore the importance of systematically exploring different model specifications to assess the robustness of inferences.

SimDE app: Simulating and visualizing formal theories using differential equations

Friday, 19th July - 09:30: Longitudinal Designs and Methods (NB C) - Oral

Rohit Batra (University of California, Davis), Meng Chen (University of Oklahoma Health Sciences Center), Emorie Beck (University of California, Davis), Emilio Ferrer (University of California, Davis)

Psychological theories are often expressed verbally using natural language, leading to varying interpretations of the phenomenon under study. We can mitigate this potential confusion by formalizing verbal theories using mathematical languages, which can help in comparing, analyzing, and interpreting one's hypotheses in quantitative terms. Differential equations (DE) are aligned with many dynamic theories in psychology. However, there is a lack of tools that can aid in the translation of verbal theories into DE systems. To facilitate this translation, we introduce *SimDE*, an open access R Shiny application that allows users to specify a DE model for a multivariate system of variables and then simulate the trajectories of each variable over time.

The users will be able to simulate a range of DE models, with features such as: 1) first- or second-order differential equations (patterns like exponential, oscillatory etc.), 2) models with or without a dynamic error term (ordinary or stochastic differential equations), 3) models with multivariate effects (coupling dynamics). The users will have the flexibility of plotting these systems and their phase space in order to see the pattern of changes over time and determine the appropriateness of the model for the phenomenon they are trying to simulate. The goal of our app is to serve as a tool for researchers who want to explore DE models for their psychological verbal theories before they even collect their data. It can also help researchers to study the implicit assumptions of systems defined with such DEs and further refine their formal models.

Integral dimensionality reduction in models that include nonlinear random coefficients

Friday, 19th July - 09:45: Longitudinal Designs and Methods (NB C) - Oral

Dr. Shelley Blozis (University of California, Davis), Dr. Jeffrey Harring (University of Maryland)

Nonlinear mixed-effects models provide a framework to study variables that follow some form of nonlinear change, such as performance measures on a learning task for which subjects show gradual improvement in performance, or variables characterized by individual differences in the within-subject variability across time, such as daily affect measures that show individual differences in the degree of variability about the individual's mean affect across days. For variables that change in a nonlinear way, nonlinear mixed-effects models permit the coefficients of a nonlinear growth function to vary between subjects. For variables that show between-subject heterogeneity in the within-subject residual variance, nonlinear mixed-effects models permit modeling of the residual variance to include a random scale effect. Both types of problems are computationally demanding due to the inclusion of a nonlinear random coefficient because the marginal distribution requires an integral with dimensions equal to the total number of random coefficients, including the linear coefficients. One way to simplify the problem is to analytically remove the nonlinear coefficient from the marginal response distribution. This approach has been applied to problems in which there is only one nonlinear coefficient, but interesting problems arise in practice where more than one nonlinear coefficient is needed. Without a reduction in the dimensions of the integral, the problem is highly demanding computationally. This talk describes integral dimensionality reduction in models that include more than one nonlinear random coefficient. An example is provided.

New Insights in Longitudinal Data and Mediation Analysis

Friday, 19th July - 09:00: Symposium: New Insights in Longitudinal Data and Mediation Analysis (NB D) - Symposium Overview

Prof. Zhiyong Zhang (University of Notre Dame)

This symposium consists of five talks that will discuss some new findings in longitudinal data analysis and mediation analysis. The first talk by Lijuan Wang will compare the use of the one-step vs. two-step approaches to modeling dynamic components in intensive longitudinal data analysis. The second talk by Xin Tong will propose a Bayesian method to handle missing data in quantile growth curve modeling of longitudinal data. The third talk by Ke-Hai Yuan will investigate the direction change of mediation effects in manifest vs. latent variable modeling. The fourth talk by Zhiyong Zhang will evaluate the use of word embedding for mediation analysis with text data where text data serve as a mediator. These presentations collectively promise to illuminate new pathways in the analysis of longitudinal and mediation data. The fifth talk by Yong Wen will focus an application of longitudinal study of the impact of living style on the health and mortality risk of the elderly.

Modeling dynamic components in intensive longitudinal data analysis: A comparison of one-step vs. two-step approaches

Friday, 19th July - 09:00: Symposium: New Insights in Longitudinal Data and Mediation Analysis (NB D) - Symposia

Dr. Peggy Wang (University of Notre Dame), Ms. Yuan Fang (University of Notre Dame)

Intensive longitudinal data are more and more frequently collected in psychology to help understand intraindividual psychological processes and interindividual differences in the dynamic processes. In this talk, we will evaluate advantages and limitations of one-step and two-step approaches of modeling dynamic components with intensive longitudinal data. In the two-step approaches, dynamic component estimates will be extracted from a dynamic model (e.g., a multilevel autoregressive model) in the first step and modeled as inputs, mediators, or outcomes in the second step. In the one-step approach, the dynamic model and the model of the relations between dynamic components and other variables are estimated simultaneously. Simulation results will be presented to compare the performance of the approaches. Implications of the simulation results and future research directions will be discussed.

Quantile Longitudinal Data Modeling with Missing Data

Friday, 19th July - 09:00: Symposium: New Insights in Longitudinal Data and Mediation Analysis (NB D) - Symposia

Dr. Xin Tong (University of Virginia), Dandan Tang (University of Virginia)

Longitudinal research often faces methodological challenges. In this study, we introduce a robust Bayesian approach using conditional quantiles to address the nonnormality of data and population heterogeneity challenges in longitudinal studies. By converting the problem of estimating a quantile longitudinal model into a problem of obtaining the maximum likelihood estimator for a modified model with the assistance of the asymmetric Laplace distribution, Bayesian estimation methods can be conveniently used. Ignorable and non-ignorable missing data will be accounted for using multiple imputation and a selection model approach, respectively. Simulation studies have been conducted to evaluate the numerical performance of the quantile approach. A real data example will also be presented in the talk to illustrate the application of the robust Bayesian quantile growth curve modeling.

Analysis of Mediation Direction Change Between Manifest and Latent Variable Modeling

Friday, 19th July - 09:00: Symposium: New Insights in Longitudinal Data and Mediation Analysis (NB D) - Symposia

Dr. Ke-Hai Yuan (University of Notre Dame)

Mediation analysis plays an important role in understanding causal processes and effects of intervention. The analysis can be conveniently conducted using least-squares (LS) regression with composite scores and followed by a significance statistical test on the indirect effect. However, measurements in social and behavioral sciences typically contain errors, and parameter estimates of LS regression are affected by such measurement errors. In particular, unreliable measurements not only affect the size but also cause sign changes of the direct and indirect effects. Using statistical learning and analysis, the current article studies what factors contribute to the sign change of the direct and indirect effects between latent variables modeling and LS regression. Parameters under the latent variable model that might contribute to the sign change are examined analytically. They are further verified numerically through many conditions on the population values of the parameters. These conditions are further used to identify the causes for sign changes of parameters between latent-variable model and LS regression using composite scores. The findings not only advance the understanding of the results of mediation analysis with composites but also provide the basis for model diagnostics to avoid misinterpreting the results in empirical studies.

Mediation Analysis with Text Data

Friday, 19th July - 09:00: Symposium: New Insights in Longitudinal Data and Mediation Analysis (NB D) - Symposia

Prof. Zhiyong Zhang (University of Notre Dame)

Qualitative text data are widely collected in research and can come from many different sources. For example, in diary studies, daily records on the activities and feelings of a day can be collected from students and/or teachers (Oppenheim, 2000). Text data can also come from the transcription of audio and video conversations from class observations (Bailey, 2008). For data collection using surveys or questionnaires, free response items (open-ended questions) are frequently used to solicit feedback (Rohrer, Brümmer, Schmukle, Goebel, & Wagner, 2017). For example, in teaching evaluation, students are often asked to elaborate on what can be improved. The responses are typically in the free text format. Compared to quantitative data collected through Likert scales, text data can provide more subtle information regarding teaching. However, text data are largely under-analyzed in social, behavioral and education research. In this study, we present a model that can treat the text data as a mediator to understand how text information can be extracted to explain the association between input variables and output variables. A two-stage method will be used to first extract the information from the text data through sentence encoders and then the information can be used in a mediation model in the structural equation modeling framework. We show how to conduct the analysis using an R package BigSEM that we develop.

A Longitudinal Study of the Impact of Living Style on the Health and Mortality Risk of the Elderly

Friday, 19th July - 09:00: Symposium: New Insights in Longitudinal Data and Mediation Analysis (NB D) - Symposia

Prof. YONG WEN (Nanjing University of Posts and Telecommunications), Mr. ZHOUQIANG YU (Nanjing University of Posts and Telecommunications)

Healthy aging and mortality are global concerns of elderly. This study investigates how living style affects the health and mortality of elderly under the typical Chinese culture, based on a longitudinal survey from 2014 to 2018. Using the panel data models and Cox proportional hazards model, mechanisms that directly affect the health and mortality rates of elderly were clarified. Results indicate that 1) living with children has a significant positive impact on the psychological health of elderly but a significant negative impact on their physical health; 2) path analysis indicates that social activities, physical exercise as well as better economic conditions contribute positively to the health status of the elderly; 3) heterogeneity analysis showed that for elderly who lost their spouses, staying with children is significantly more beneficial for those who are in the rural area than the city dwellers. In contrast, the benefit of living with children for couples is significantly less. 4) Living style significantly affects the mortality risk of elderly, especially for those over 80 and in the rural area.

Constructing Language Models to Predict Massive Student Sentiment Toward Universities

Friday, 19th July - 09:00: Classification Methods (RB 209) - Oral

Tonghui Xu (University of Massachusetts Lowell), Prof. Xiaobai Li (University of Massachusetts Lowell), Prof. Hsien-Yuan Hsu (University of Massachusetts Lowell), Dr. Yan Wang (University of Massachusetts Lowell)

The use of language models (LMs) to understand sentiment in extensive open-ended questions has been introduced in educational and psychological studies. However, most related LMs studies focus on student reviews of faculty or courses, overlooking student satisfaction with university features (e.g., safety and academics). Additionally, these open-ended questions contain diverse information and complex sentiments. Analyzing the sentiment of large sample data may require significant effort. Consequently, these questions may need to be ignored, potentially leading to the loss of valuable information richness. While utilizing ChatGPT for analysis is an approach, there remains a risk of data disclosure. To address these limitations, we propose constructing a specific LM to comprehend student sentiments toward universities by Python language. We collected over 200,000 online student reviews and utilized 75% of the data for training four neural network models with 5-fold cross-validation, each with 10 epochs: (a) Long Short-Term Memory (LSTM), (b) Convolutional Neural Network (CNN) with word2Vec, (c) CNN+LSTM, and (d) LSTM+CNN. The remaining 25% was used for testing. We compared these models to determine the optimal one and saved it. The LSTM+CNN model demonstrates strong training and test performance, achieving an overall classified accuracy of 85% and 78% and AUC values of 95% and 86%, respectively. Positive sentiment precision and recall stand at 85% and 80%, while negative sentiment precision and recall are 71% and 78%. These results indicate a low risk of overfitting and effective identification of sentiment in student review data, even when dealing with new data.

Advancing enemy item detection using transformer-augmented NLP

Friday, 19th July - 09:15: Classification Methods (RB 209) - Oral

Dr. Paulius Satkus (Graduate Management Admission Council), Dr. Yan Fu (Graduate Management Admission Council), Dr. Kyung (Chris) Han (Graduate Management Admission Council)

The identification of enemy items—items that share the same or extremely similar contents and/or that potentially offer clues to other items —presents significant operational challenges due to the sheer number of unique item pairs within an item pool. Although previous researchers (Becker & Kao, 2022; Fu & Han, 2022; Micir et al., 2022) have demonstrated the utility of Natural Language Processing (NLP) methods in narrowing down potential enemy item lists, their approaches predominantly relied on simpler textual analysis techniques. This research aims to enhance the process of identifying these items by transitioning from string-based and corpus-based approaches to the leading edge of NLP technology, specifically, transformer-based neural network models, which offer advancements in capturing deep contextual relationships within text (Vaswani et al., 2017).

To predict item enemy status, we utilized items from a high-stakes testing program to develop an extreme gradient-boosting machine learning model (Chen & Guestrin, 2016), which has been effective in classification tasks with imbalanced data (Hancock et al., 2023). The predictors were the item pair similarity indices based on strings (e.g., N-gram, ROUGE), TF-IDF, latent semantic analysis, and item embeddings (from BERT family models and GPT2). Due to highly imbalanced data, we used F1 scores to evaluate model performance. The results are encouraging: F1 scores ranged from .63 to .80 for overall and content-domain models. These findings indicate that integrating newer NLP methods can be useful for accurately identifying high-similarity item pairs and thereby streamlining the review workload for content experts more effectively than earlier approaches.

Improving Context Scale Interpretation Using Latent Class Analysis for Cut Scores

Friday, 19th July - 09:30: Classification Methods (RB 209) - Oral

Dr. Liqun Yin (Boston College (TIMSS & PIRLS International Study Center)), Dr. Ummugul Bezirhan (Boston College (TIMSS & PIRLS International Study Center)), Dr. Matthias Von Davier (Boston College (TIMSS & PIRLS International Study Center))

This paper introduces an approach that uses latent class analysis to identify cut scores (LCA-CS) and categorize respondents based on context scales derived from large-scale assessments like PIRLS, TIMSS, and NEAP. Context scales use sets of Likert scale items to measure latent constructs of interest. For interpretation, respondents are classified into high, middle, and low regions utilizing specified cut-points on the context scale. Unlike conventional methods reliant on human judgments to define cut-points based on item content, latent class analysis is a categorical latent variable modeling technique that allows identifying groups according to a statistical optimality criterion. LCA finds a categorical latent variable that explains differences in item scores based on score distribution differences between homogeneous groups (latent classes). To derive cut-points using this approach, classes are sorted, and conditional score distributions given class are smoothed under the assumption that each class is a homogeneous group with a conditional normal distribution of the scores given class. This allows us to derive the conditional probability of class membership, which provides the basis for finding cut scores. Finally, cut-points are identified by locating the intersection point of adjacent smoothed posterior probability distributions and connecting it to the construct. Demonstrated through application to PIRLS 2021 data, this method not only validates existing categorizations on the context scales but also enhances classification accuracy, particularly for scales exhibiting highly skewed distributions across diverse countries. Recommendations for researchers to adopt this LCA-CS approach are provided, demonstrating its efficiency and objectivity compared to traditional judgment-based approaches.

Blueprinting the Future: Automatic Item Categorization using Hierarchical Zero-Shot and Few-Shot Classifiers

Friday, 19th July - 09:45: Classification Methods (RB 209) - Oral

Dr. Ting Wang (American Board of Family Medicine)

In testing industry, precise item categorization is pivotal to align exam questions with the designated content domains outlined in the assessment blueprint. Traditional methods either entail manual classification, which is laborious and error-prone, or utilize machine learning requiring extensive training data, often leading to model underfit or overfit issues. This study unveils a novel approach employing the zero-shot and few-shot Generative Pretrained Transformer (GPT) classifier for hierarchical item categorization, minimizing the necessity for training data, and instead, leveraging human-like language descriptions to define categories. Through a structured python dictionary, the hierarchical nature of examination blueprints is navigated seamlessly, allowing for a tiered classification of items across multiple levels. An initial simulation with artificial data demonstrates the efficacy of this method, achieving an average accuracy of 92.91% measured by the F1 score. This method was further applied to real exam items from the 2022 In-Training Examination (ITE) conducted by the American Board of Family Medicine (ABFM), reclassifying 200 items according to a newly formulated blueprint swiftly in 15 minutes, a task that traditionally could span several days among editors and physicians. This innovative approach not only drastically cuts down classification time but also ensures a consistent, principle-driven categorization, minimizing human biases and discrepancies. The ability to refine classifications by adjusting definitions adds to its robustness and sustainability.

Multilevel synthesis of composite cross-classified event scores as a psychometric approach for EdTech

Friday, 19th July - 09:00: Psychometric Applications to Education (RB 210) - Oral

Dr. Dmitry Abbakumov (Cognitio)

Learner behavior in eLearning settings generates complex data logs, encompassing attempts, response times, interactions with hints, GPT-based support, and various other elements. The diversity of this logged information complicates modeling learner and content properties using traditional Item Response Theory (IRT) methodologies, typically not designed to directly incorporate such a wide range of interaction data. To address this challenge, we introduce a novel approach consisting of two stages. Initially, we perform composite event scoring, integrating all logged information into a singular, cross-classified score, x_{ij}. This method utilizes cross-product attempt tiering, a novel technique, to ensure score comparability across a range of educational products, including courses and disciplines. Subsequently, we apply a multilevel synthesis technique to derive measures for the learner (e.g., individual and cohort performance) and the content (e.g., item, lesson, topic, and course feasibility). The validity of these measures is demonstrated by their correlation with conventional IRT parameters in both simulated and real-data settings. Furthermore, we explore the relationship between these measures and key EdTech business metrics, including conversion rates (CR), retention rates (RR), and completion rates (CoR). This exploration enables businesses to optimize learner experiences and content quality in accordance with the provided measures, resulting in added business profit. The efficacy of our approach is evidenced by three detailed business case studies. This work extends our research trajectory on analyzing eLearning data, previously presented at the IMPS conferences in 2018, 2019, and 2020, demonstrating the pivotal role of psychometrics in the EdTech industry.

Latent logistic regression analysis of background characteristics on digital literacy

Friday, 19th July - 09:15: Psychometric Applications to Education (RB 210) - Oral

Dr. Qianru Liang (Jinan University), Prof. Jimmy de la Torre (The University of Hong Kong), Prof. Nancy Law (The University of Hong Kong)

This study examines the relationship between students' gender and socioeconomic status (SES) and their digital literacy subskills. Data were collected from three age cohorts (Primary 3, Secondary 1, and Secondary 3) in Hong Kong and analyzed using latent logistic regression based on a corrected three-step approach of cognitive diagnosis modeling with covariates. This study also compared the three cohorts to determine if the relationships vary at different ages. We found that gender differences in favor of females were only present at Secondary 1 and 3. At Secondary 1, girls outperformed boys on all subskills, while at Secondary 3, significant gender differences were only found in one subskill, namely, communication and collaboration. Additionally, we found a positive correlation between SES and all subskills across all three age groups. When gender and SES were put in the models simultaneously, the interactions between gender and SES were only significant at Secondary 1, but not in the other two cohorts. Significant gender differences and interactions were observed in all subskills at Secondary 1. The multiple logistic regression results for Primary 3 were similar to their simple latent logistic regression results, whereas the results for Secondary 3 showed that when gender was in the model, SES was no longer predictive with respect to some subskills. Further longitudinal research is needed to better understand how the relations between DL skills and gender or SES change over time.

Accounting for ceiling effects on polytomous responses. An illustration with gender equality endorsement.

Friday, 19th July - 09:30: Psychometric Applications to Education (RB 210) - Oral

Dr. Diego Carrasco (Centro de Medición Mide UC, Pontificia Universidad Católica de Chile), Dr. David Torres Irribarra (Escuela de Psicología, Pontificia Universidad Católica de Chile)

Gender equality endorsement is an intergroup measure present in different survey-based studies and is a prominent measure among sustainable developmental goals (SDG) (Sandoval-Hernández, & Carrasco, 2020). To this end, countries can rely on the gender equality endorsement scale included in the International Civic and Citizenship Study (ICCS), which provides probabilistic samples of 8th-grade students from different countries and assesses gender equality endorsement between men and women.

Traditional methods to generate scores on this scale rely on the partial credit model (PCM), a response model that uses a normally distributed latent variable to represent students' propensity to respond to the different included items. Moreover, researchers rely on regression-based methods to inquire about related factors and program evaluation effects. However, the responses to this collection of items are highly skewed. This skewness is desirable. It means a noticeable portion of students endorse gender equality at the scale ceiling. Nevertheless, traditional regression models may produce distorted estimates in the presence of ceiling effects on the response variable (e.g., Lidell & Kruschke, 2018; Šimkovic & Träuble, 2019).

We propose a method that relies on the property of monotonicity of the PCM scores, and create a reverse sum total score. We use zero-inflated models to separate ceiling cases from the rest of the scores, allowing us to make inferences on both sides: the students at the ceiling and those in the rest of the distribution. We argue this method is a useful tool for program evaluations dealing with ceiling effects in their attribute of interest.

Score-Difference-at-Risk (SDaR): Using risk metrics to anticipate negative impact.

Friday, 19th July - 09:45: Psychometric Applications to Education (RB 210) - Oral

Dr. Sergio Araneda (Caveon)

In this presentation, I will introduce the concept of “Score Difference at Risk (SDaR),” a metric designed to quantify the risk of extreme cases of score difference due to estimation error. SDaR is inspired by the concept of Value-at-Risk (VaR) used in finances. For this metric, you calculate the value in some distribution tail where you attain a certain probability maximum. By calculating those limit numbers, you can check if the probability of some extreme event overpasses the maximum probability level you specified in advance. This metric aims to mitigate the limitations inherent in traditional variance-based statistics within psychometrics by providing a metric that connects in a better way potential negative consequences with error of measurement.

I will explain how you calculate the metric using simulations and IRT score estimations, and a series of practical examples where this metric can be used: Multiple-Stage Testing, Item parameter drift, and Test Security. I will show the calculations of the SDaR Metric for both sides of the score distribution, and also a percentage of examinees per bucket in the theta scale that are considered “at-risk” using some arbitrary criteria to illustrate the use of this new metric. I will show how this metric can be used as an alternative way to evaluate conditional errors of measurement, and how it can also be used to see marginal impacts on error of measurement due to different decisions about a test.

An Interpretable Cognitive Diagnostic Model Based on XGBoost and Shap

Friday, 19th July - 09:00: Topics in Machine Learning (RB 211) - Oral

Ms. Chang Nie (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China), Prof. Tao Xin (Beijing Normal University)

Cognitive diagnosis, which is used to assess students' cognitive abilities, is a widespread concern in educational science. At present, researchers around the world have developed at least 100 kinds of cognitive diagnostic models, including cognitive diagnostic models based on traditional probabilistic statistical methods and cognitive diagnostic models based on machine learning algorithms. It has been proved that machine learning-based models have higher flexibility and accuracy, and have obvious superiority in small-sample situations. However, the "black box" nature of such models leads to poor interpretability, which is not easy to be trusted in real-world teaching practice. Therefore, we attempt to construct an interpretable machine learning-based cognitive diagnosis model. First, we construct a cognitive diagnostic model based on the XGBoost algorithm and test its accuracy under different test conditions. Subsequently, we illustrate and present the interpretability of the new model with Shap values as well as some interesting visualisation results.

Understanding of Measurement Invariance within a SEMtree Framework

Friday, 19th July - 09:15: Topics in Machine Learning (RB 211) - Oral

Mr. YeongJin Jo (Yonsei University), Prof. Ji Hoon Ryoo (Yonsei University)

Measurement invariance (MI) plays a key role in multi-group analysis and longitudinal studies in social and behavioral sciences. The necessity of confirming the MI in such designs has been emphasized over the last three decades. On the other hand, such an importance has been de-emphasized when dealing with big data such as SEMtree. Brandmaier et al (2014) and Finch (2017) pointed out measurement invariance within the SEMtree framework. However, it has not been proposed how to examine the MI and what to be considered when interpreting the results. This study explores the **concept of the MI within the SEMtree framework**, proposes several tests to **confirm the MI over heterogeneous groups** defined from the decision tree, and **demonstrates examples** of conducting the MI to show how to interpret the results correctly.

- The first task is to identify three types of tests: Split-wise MI tests introduced by Brandmaier et al, specific tests via pairwise MI tests over final nodes, and omnibus test utilizing alignment method.

SEMtree aims to cluster homogeneous groups as final nodes, and to examine the nonlinear interactions between covariates when splitting nodes. In the process, there are many decision makings such as considering criteria of splitting, which means that the final nodes would be best-fitting ones but may not be the only one, i.e., the splitting itself is a random step. Thus, the goal of the MI should be sensitivity and stability instead of a deterministic sense.

- The second task is to confirm those tests via a simulation study.

Using Conditional Tabular Generative Adversarial Networks for Process Data Generation in Monte Carlo Simulation Studies

Friday, 19th July - 09:30: Topics in Machine Learning (RB 211) - Oral

Dr. Youmi Suk (Teachers College, Columbia University), Dr. Ke Yang (University of Texas at San Antonio)

Monte Carlo simulation studies are commonly employed to evaluate both classical and modern psychometric methods. In traditional simulation studies with response process data, researchers often generate data with a high degree of smoothness and limited interdependencies among variables, diverging from real-world complexities. Given researchers' discretion in data generation, the performance of methods in such settings may often fail to accurately reflect their efficacy in real-world scenarios. To enhance the credibility of such simulations, we propose using Conditional Tabular Generative Adversarial Networks (CTGANs), a class of generative AI models designed for tabular data generation, to generate realistic process data. We present enhanced algorithms tailored for CTGANs to generate tabular process data. Using the enhanced CTGANs, we then outline a specific procedure for conducting AI-based simulation studies aimed at evaluating the predictive performance of psychometric methods. Notably, we demonstrate the proposed AI-based simulation approach by evaluating the performance of Tang et al.'s (2020) method, which utilizes multidimensional scaling to extract latent features from action sequences in process data. Lastly, we discuss general guidelines for conducting AI-based simulation studies to evaluate predictive performance.

Unveiling competencies for efficient AI utilization: A psychometric perspective

Friday, 19th July - 09:45: Topics in Machine Learning (RB 211) - Oral

Ms. Bhavana Srirangam (Mercer | Mettl), Ms. Karishma Agarwal (Mercer | Mettl), Dr. Rob Bailey (Mercer | Mettl)

The emergence of Generative Artificial Intelligence (Gen AI) tools has transcended technical boundaries, transforming AI into an invaluable tool for enhanced performance and productivity across industries. For organizations to successfully adopt AI, 60% of the current workforce will need AI upskilling/reskilling (Oliver Wyman, 2024). Consequently, researchers have highlighted the need to identify and measure key competencies for effectively working with AI. While past literature has prioritized technical aptitude for working with AI, there is a paucity of research examining the crucial role of personality in AI adoption and utilization.

The present study aims to identify personality-based competencies contributing to efficient AI utilization. Based on an extensive review of existing academic literature and industry trends, relevant competencies were identified and a framework was developed. The framework competencies include adaptability (*openness and adaptability to change*), innovation (*generating and implementing novel solutions*), change management (*embracing and advocating change*) and diversity and inclusion (*fostering inclusive collaboration for collective success*).

The competencies were assessed using the Mettl Personality Map (MPM). To validate the framework, confirmatory factor analysis was conducted with a heterogenous global sample of working professionals (n=7001; average age=34.04 ± 10.17 years). Results affirm the construct validity and factorial validity of the competencies. For all competencies, the trait factor loadings exceeded .50 (p<.001). Also, each competency exhibited an acceptable model fit (CFI>.95, TLI>.95; RMSEA<.08, SRMR<.05; composite reliability>.70). By assessing relevant competencies, the framework can benefit organizations in catalyzing AI-adoption.

Rating systems for measurement in adaptive learning systems: Challenges and solutions

Friday, 19th July - 10:30: Invited Talk (Vencovského aula) - Invited Talk

Dr. Maria Bolsinova (Tilburg University)

Adaptive learning systems (ALS) are a branch of technology-enhanced learning platforms that optimize the learning and practice material based on the student's recent activity and provide feedback on their educational progress over time. To optimize feedback and learning material, one needs to have continuously updated, accurate, and reliable measures of the students' changing abilities. This makes measurement of change one of the central issues in ALS. Measuring change is made even more challenging because of the adaptive nature of ALS and because they operate at a large scale with thousands of students needing continuous updates of their ability estimates. A possible solution to these challenges is provided by the Elo Rating System (ERS) which allows one to track the development of student abilities and item difficulties by updating their corresponding ratings after every response. The measurement properties of ERS are not well known, and so in this presentation, I will first present results on the properties of Elo ratings: 1) the estimates are generally biased, 2) the variances are context-dependent and not known in advance, and 3) when the items are presented to students adaptively, bias cannot be defined because the ratings drift away from each other after every response. Second, I will introduce a modification of Elo which solves the issue of ratings drifting away from each other. Third, I will present a new urn-based rating system called Urnings which solves not only the drift issue, but also the problem of bias and unknown variance.

Using robust scaling to address DIF and DTF in latent variable models

Friday, 19th July - 10:30: Invited Talk (RB 101) - Invited Talk

Dr. Halpin, Peter Francis (University of North Carolina at Chapel Hill)

The overall argument of this talk is that IRT-based scaling and differential item functioning (DIF) are two sides of the same problem. In particular, DIF with respect to a grouping variable is formally similar to IRT-based scaling with the common items non-equivalent groups (CINEG) design. Items with DIF translate into outliers in the CINEG design, and this outlier detection problem is remarkably amenable to existing methods from robust statistics. The utility of this overall approach is illustrated in two contexts. First, I show how robust scaling can be used to construct a DIF-detection procedure that (a) does not require pre-specification of anchor items, (b) comes with theoretical guarantees about its performance when fewer than 1/2 of items exhibit DIF, and (c) can be conveniently applied as a post-estimation procedure following separate calibrations (configured invariance) of a wide class of unidimensional latent variable models. Second, I show how robust scaling yields a Hausman-like specification test of whether DIF affects group comparisons on the latent trait (i.e. differential test functioning or DTF). The test does not require identifying which, if any, items may exhibit DIF, thereby obviating the need for item-by-item analyses before evaluating DTF. I illustrate the usefulness of the specification test for addressing concerns about the quality of outcome measures used in education research, focusing in particular on the interpretation of DTF with respect to randomly assigned treatment conditions. Finally, I discuss ongoing work to extend robust scaling to multiple groups, longitudinal settings, and multidimensional models.

User Beware - Emerging Evidence that Survey Item Calibration and Scoring Decisions Can Introduce Substantial Bias into Growth Mixture Model Results

Friday, 19th July - 11:30: Symposium: User Beware - Emerging Evidence that Survey Item Calibration and Scoring Decisions Can Introduce Substantial Bias into Growth Mixture Model Results (RB 101) - Symposium Overview

Dr. Veronica Cole (Wake Forest University)

Growth mixture models (GMM) have long been used to find homogeneous groupings of individuals based on growth trajectories. However, there is ample evidence that small changes in data processing or model specification impact GMM results. Seemingly trivial decisions drastically alter the number and nature of latent classes. One set of decisions involves scoring repeated measures that are the foundation of a GMM. There are numerous ways to produce scores used in GMMs; from simple sum scores all the way to multidimensional item response theory (IRT) models with different model parameters at different time pointsThe presentations in this symposium seek to answer one question: how do decisions made in the scoring process affect the fit and interpretation of the GMMs ? A fundamental tension arises because an IRT model that does not account for group membership often assumes the latent variable means and variances come from a single population. Yet, when producing scores for GMMs, one cannot fit a multigroup IRT model because the classes are latent. If data truly arise from a from an unobserved grouping, most scoring models will essentially be misspecified. This problem is compounded by several other issues which arise in scoring, such as the question of whether unmodeled response styles can lead to spurious classes, and how covariates used in the scoring process should be incorporated into the GMM itself. We tackle these separate but related issues in this symposium, drawing on theory as well as results from analyses of real data and Monte Carlo simulations.

Latent trajectory or measurement artifact? Understanding the effects of unmodeled response styles on growth mixture modeling results.

Friday, 19th July - 11:30: Symposium: User Beware - Emerging Evidence that Survey Item Calibration and Scoring Decisions Can Introduce Substantial Bias into Growth Mixture Model Results (RB 101) - Symposia

Dr. Veronica Cole (Wake Forest University), Dr. James Soland (University of Virginia), Stephen Tavares (University of Virginia), Qilin Zhang (Wake Forest University)

Growth mixture models (GMM) aim to find homogenous groups of individuals based on trajectories of growth over time. The inputs to these models are often scores from a self-report measure assessed at multiple time-points. While these scores ideally represent some latent phenomenon which changes over time, scores are obviously a function of many things, for example bias from response styles (e.g., the tendency to use extremes of the response scale in ways unrelated to the construct of interest). This project uses simulation to examine whether unmodeled response style bias may manifest as spurious classes in GMM results. Data are generated from a single-class model including a latent factor which changes over four timepoints, as well as four factors representing acquiescent responding[JS1] over time. Scores are then generated from these items using methods which do not take response style bias into account, and those that do account for response style bias, including the multidimensional nominal response model (MNRM). GMM's with varying numbers of classes are then fit to these scores. Our interest lies in whether fit indices erroneously favor GMM's with more than one class when the response style factor is not modeled. Initial results indicate that ignoring response style bias when modeling the item responses at a single timepoint can lead to an inappropriate number of classes being extracted. Results presented at the conference will include GMM simulations, as well as results from empirical socio-emotional learning data based on surveys administered to nearly 1.5 million students in California.

Why Producing Unbiased Survey Scores for Use in Growth Mixture Models is Almost Impossible: A Few Illustrations

Friday, 19th July - 11:30: Symposium: User Beware - Emerging Evidence that Survey Item Calibration and Scoring Decisions Can Introduce Substantial Bias into Growth Mixture Model Results (RB 101) - Symposia

Dr. James Soland (University of Virginia), Stephen Tavares (University of Virginia), Dr. Veronica Cole (Wake Forest University), Qilin Zhang (Wake Forest University)

Interest in identifying latent growth profiles to support the psychological and social-emotional development of individuals has translated into the widespread use of growth mixture models (GMMs). In most cases, GMMs are based on scores from item responses collected using survey scales or other measures. Research already shows that GMMs can be sensitive to departures from ideal modeling conditions, and that growth model results outside of GMMs are sensitive to decisions about how item responses are scored, but the impact of scoring decisions on GMMs has never been investigated. We start to close that gap in the literature with the current study. Through empirical and Monte Carlo studies, we show that GMM results—including convergence, class enumeration, and latent growth trajectories within class—are extremely sensitive to seemingly arcane measurement decisions. Further, our results make clear that, because GMM latent classes are not known a priori, measurement models used to produce scores for use in GMMs are, almost by definition, misspecified because they cannot account for group membership. Misspecification of the measurement model then, in turn, biases GMM results. Practical implications of these results are discussed. Our findings raise serious concerns that many results in the current GMM literature may be driven, in part or whole, by measurement artifacts rather than substantive differences in developmental trends.

Investigating the Sensitivity of Growth Mixture Model Results to Conditioning Scoring Decisions on Time-Invariant Covariates

Friday, 19th July - 11:30: Symposium: User Beware - Emerging Evidence that Survey Item Calibration and Scoring Decisions Can Introduce Substantial Bias into Growth Mixture Model Results (RB 101) - Symposia

Stephen Tavares (University of Virginia), Dr. James Soland (University of Virginia), Dr. Veronica Cole (Wake Forest University)

Growth Mixture Models (GMMs) offer a pathway toward a person-centered perspective in research involving growth and development over time. The question of when, where, and how to introduce covariates into GMM models continues to be a topic of ongoing research. Previous research has shown that conditioning latent factors on covariates can impact the results of second-stage analyses, including factor mixture models (FMMs) and subsequent regressions that incorporate latent factors. While covariate impact has been assessed on manifest variables, latent growth parameters, and latent class variables, there is a need to investigate the impact of a covariate if an IRT approach is used to score survey instruments. This paper investigates the direct effects of a covariate on latent IRT scores through simulations and an empirical example. For our simulations, data are generated from a six-class model where three classes represent one known class where our covariate is 0, and the other three classes represent the other known class where our covariate is 1. Scores are then generated using various IRT scoring methods and under conditions where the scores are conditioned on the covariate. GMs are subsequently assessed across different classes. Fit indices are then leveraged to determine optimal class enumeration as well as the modeled growth trajectory associated with each class. Initial results show that conditioning latent scores on a covariate enhances class enumeration only marginally. Results presented include the results of the simulations as well as the results from 5,000 Australian children.

Comparing trajectories of school-to-work transitions using Latent Transition Analysis

Friday, 19th July - 11:30: Applications of Longitudinal Data Analysis (NB A) - Oral

Dr. Tomasz Żółtak (Educational Research Institute), Dr. Grzegorz Piotr Humenny (Educational Research Institute)

In the presentation, I will compare the differences in the trajectories of school-to-work transitions of Polish secondary schools graduates of two cohorts: those graduating in 2020, i.e. few months before the first peak of the SARS-CoV-2 epidemic in Poland, and in 2021, when the epidemic started to subside. To this end, I will use Latent Transition Analysis (LTA) applied to a unique dataset covering all Polish secondary schools' graduates from 2020 and 2021, along with information on their educational and occupational status in the twenty months after graduation, obtained from the Polish central education and social security registers.

The presentation will discuss the methodological and technical aspects of applying LTA and its variant, including "random intercepts" (RI-LTA: Muthén & Asparouhov, 2022) to large datasets covering long periods with relatively high temporal granularity. Also, building on the flexibility of the LTA approach, I will describe how various characteristics of graduates (educational track: gender, obtaining formal certificates, graduating in professions most affected by pandemic restrictions) and contextual factors (local economy and labor market indicators) shaped the probability of following different trajectories of the school-to-work transition. Finally, I will discuss some limitations of the analysis, resulting primarily from using registry data – covering a large population but providing only a limited set of characteristics.

Identifying subject heterogeneity in longitudinal brain connectivity patterns with Clusterwise-IVA

Friday, 19th July - 11:45: Applications of Longitudinal Data Analysis (NB A) - Oral

Dr. Tom Wilderjans (Leiden University), Mr. Jeffrey Durieux (Erasmus University Rotterdam)

Nowadays, in different fields of science (e.g., neuroscience and psychology) researchers are interested in uncovering the processes and temporal changes therein underlying longitudinal big data. In neuroscience, for example, fMRI brain scans are taken from patients at regular moments in time (e.g., yearly sessions) in order to identify temporal changes in functional connectivity (FC) patterns (i.e., correlated brain regions collaborating in psychological functioning) related to a particular disease. Previous research in this regard revealed longitudinal changes in FC patterns that are typical for dementia patients (Dautricourt et al., 2021). To extract temporal changes in FC patterns from longitudinal multi-subject fMRI data, Independent Vector Analysis (IVA) is proposed, which performs ICA on the data of each session and assures that associated components are dependent across sessions. As such, the extracted components capture longitudinal changes in FC patterns. When studying brain diseases (e.g., dementia and depression), however, often sample heterogeneity in the underlying processes and in the temporal changes therein exists as brain diseases often develop in heterogeneous ways across patients (groups). Uncovering this heterogeneity is possible by clustering subjects based on differences in temporal change profiles in FC patterns. To this end, we propose Clusterwise IVA, which clusters subjects and at the same time estimates the longitudinally changes in FC patterns characterizing each subject cluster. In our presentation, Clusterwise IVA will be explained and the performance of the method will be evaluated by means of an extensive simulation study and/or an illustrative application to longitudinal multi-subject fMRI data from Alzheimer patients.

Ability tracking for Intelligent Tutoring Systems

Friday, 19th July - 12:00: Applications of Longitudinal Data Analysis (NB A) - Oral

Mr. Karl Sigfrid (Stockholm University)

An intelligent tutoring system (ITS) aims to provide instructions and exercises tailored to the ability of the student. To do this, the ITS needs to estimate the student ability based on student input. Rather than including frequent full-scale tests to update our ability estimate, we want to make estimates from the outcomes of practice exercises that are part of the learning process. A challenge with this is that the ability changes as the student learns, which makes traditional item response theory (IRT) models inappropriate. Most IRT models estimate an ability based on a test result, and assume that the ability is constant throughout a test.
We propose a method for measuring abilities that change throughout the measurement period. The method assumes that the abilities for a group of respondents who are all in the same stage of the learning process follow a normal distribution. The method does not assume a particular shape of the growth curve, and it is robust to rapid ability change. When we compare our method to the Elo method on chess data, our method has a performance similar to that of the Elo method when we estimate the ability based on one game per active month. However, with more time between measurements, our method performs better than the Elo algorithm. More generally, our method has an advantage when the measured outcomes are far apart in time, or when the ability changes rapidly.

Decomposing Social Disparities through Various Interventions: Incorporating Simulation-Based Sensitivity Analysis

Friday, 19th July - 12:15: Applications of Longitudinal Data Analysis (NB A) - Oral

Dr. Soojin Park (University of California Riverside)

Large disparities in cognitive, economic, and health outcomes persist across social groups in the US. Traditional approaches to decomposition (e.g., difference-in-coefficients or Oaxaca Blinder decomposition) have been utilized to identify mediators, such as educational attainment, that explain these disparities. However, these approaches have limitations, as they do not clarify assumptions (e.g., no omitted confounding) and lack a clear way to validate findings against possible violations of those assumptions. The development of causal decomposition analysis (Jackson and VanderWeele, 2018) has clarified those assumptions. Recent work on methods for disparity research (Park et al., 2023) developed a sensitivity analysis that assesses the robustness of findings against a reasonable amount of omitted confounding. Nevertheless, existing sensitivity analyses are limited, addressing only continuous outcomes and time-fixed measures.

To overcome these limitations, we extend simulation-based sensitivity analysis to address various types of mediator and outcome variables. In addition, we enahce accessibility by providing a method to benchmark the strength of unobserved confounding against observed covariates, even for binary mediators or outcomes. Compared to currently available simulation-based sensitivity analysis, our approach offers several advantages: 1) it can handle cases where unobserved confounders are dependent on existing covariates, 2) it properly addresses the uncertainly of sensitivity parameters, and 3) it can be easily extended to incorporate time varying measures. We demonstrate the effectiveness of our proposed sensitivity analysis through simulation studies and illustrate its utility in education context.

Recursive partitioning of continuous-time structural equation models

Friday, 19th July - 11:30: Topics in Structural Equation Modeling (NB B) - Oral

Dr. Pablo F. Cáncer (Universidad Pontificia Comillas), Dr. Manuel Arnold (Humboldt-Universität zu Berlin), Dr. Eduardo Estrada (Universidad Autónoma de Madrid), Dr. Manuel Voelkle (Humboldt-Universität zu Berlin)

Purpose. Model-based recursive partitioning has been gaining traction in psychological research. The technique finds similar individuals in heterogeneous data sets and identifies the most important predictors of group differences in the process. In the past decade, structural equation models (SEM) have been almost entirely partitioned using the *semtree* software package, leading to so-called SEM trees and forests. Recently, score-based covariate testing has been implemented into *semtree*, drastically improving runtime and making the partitioning of more complex models possible. In the present work, we extended this approach to continuous-time models. Unlike discrete-time (DT) models, CT models adapt effortlessly to longitudinal data observed with different time intervals between measurements. Thus, our resulting approach, which we call score-based CT-SEM trees and forests, is well suited to deal with heterogeneity between individuals and measurement occasions. However, it is uncertain whether CT-SEM forests will be feasible in terms of computation time and whether their performance will be acceptable for the empirical practice.

Method. To answer these questions, we conducted a Monte Carlo study to evaluate the performance of CT-SEM forests under a broad set of empirically relevant conditions. We also illustrated the application of a CT-SEM forest using empirical data from the Survey of Health, Ageing, and Retirement in Europe (SHARE).

Results and discussion. We discuss the most relevant findings, elaborate on the strengths and limitations of the proposed algorithm, and comment on current challenges and future lines of research in the context of between-individual differences in change.

Evaluating multigroup structural equation modeling with exploratory measurement models

Friday, 19th July - 11:45: Topics in Structural Equation Modeling (NB B) - Oral

Ms. Jennifer Dang Guay (KU Leuven), Prof. Yves Rosseel (Ghent University), Kim De Roover (KU Leuven)

Structural Equation Modelling (SEM) is the state-of-the-art for modeling relations between latent variables (e.g., anxiety and wellbeing), also called 'factors'. A SEM model consists of a measurement model, which specifies how observed indicators (e.g., questionnaire items) measure the factors, and a structural model, which captures the relations among factors. When data are available on multiple groups (e.g., cross-national data), multigroup SEM is used to compare the structural relations across groups. Recently, the Structural-After-Measurement (SAM) approach was proposed by Rosseel and Loh (2022) as an alternative way to estimate SEM. SAM estimates the measurement and structural models separately, whereas traditional SEM estimates both simultaneously. SAM improves the convergence, robustness and flexibility of SEM and allows to scrutinize the measurement model in the first step. In multigroup settings, SAM also allows to investigate whether the measurement model is invariant across groups (i.e., measurement invariance) and to identify non-invariances, before estimating and comparing the structural relations. While Confirmatory Factor Analysis (CFA) is typically used for the measurement model, recent Exploratory Factor Analysis (EFA) techniques show promise. Using the SAM approach, this study will compare several multigroup EFA techniques with different oblique rotations as measurement models in the first step. Multigroup CFA is included as a benchmark. Then, in the second step, the group-specific structural models are estimated. The aim of the study is to evaluate and compare how well the EFA-based techniques recover the measurement model and measurement (non-)invariances, and how this affects the recovery and comparison of structural relations across groups.

Small sample estimation of structural equation models: A comparison of alternative estimation approaches

Friday, 19th July - 12:00: Topics in Structural Equation Modeling (NB B) - Oral

Dr. Graham Rifenbark (University of Connecticut), Prof. Yves Rosseel (Ghent University)

Typically, structural equation model (SEM) parameters are estimated using maximum likelihood (ML) which possess desirable attributes related to point estimation and inferences; however, ML is a large sample estimation method. In practice, researchers are often confronted with small sample sizes and when ML is used in these settings a host of issues can surface, including nonconvergence, improper solutions, biased point estimates, and poor inference quality. Ultimately, latent parameters (e.g., regressions) are of most interest and the accuracy of these estimates is paramount. Alternative estimation approaches have been proposed in the context of small sample size SEM including factor score regression, the structural-after-measurement framework (SAM, Rosseel & Loh, 2022), as well as an iterative procedure proposed by Ozenne et al., (2020). We designed a Monte Carlo simulation to compare the performance of SAM and that of Ozenne et al., (2020) relative to factor score regression and traditional SEM with bounded ML (De Jonckere & Rosseel, 2022). For the SAM approach, we will evaluate two different estimation procedures at stage-1: a non-iterative estimation method (multiple group method) versus ML; and at stage-2, we will use ML. We report simulation results for various levels of construct reliability, sample sizes, and measurement model structures while specifying a population model (containing 69 free parameters) and misspecified models. We focus on structural parameter recovery and will consult bias, standard deviation, and root mean squared error.

Investigating text complexity indices in predicting SME-rated passage grade

Friday, 19th July - 11:30: AI and Machine Learning in Educational Settings (NB C) - Oral

Dr. Ann Hu (NWEA/HMH)

Selecting passages with appropriate complexity is critical in computer adaptive testing or tutoring. Text complexity can be defined by vocabulary, cohesion, genre, etc. Typically, it is quantified by the US grade level needed to comprehend the texts.

Various text statistics exist to estimate the passage grade. Understanding their accuracy could cut the cost for passage development. This study investigated eleven passage grade indices: Lexile grade derived from the measures of Lexile Analyzer (MetaMetrics, 2016), three Flesch-Kincaid Grade Levels (FKGL) from Coh-Metrix Text Analyzer (Graesser et al., 2004) and the Python libraries textstat (Bansal, 2022) and spaCy (Honnibal & Montani, 2017), plus seven additional indices from textstat. These indices were incorporated into regression models to predict the grades assigned by subject matter experts (SME) to 433 passages for grades two through twelve. Considering the high correlation among some indices (0.46 to 0.99), five machine-learning-based regression analyses were employed, including stepwise linear regression, Ridge, LASSO, ElasticNet, and Random Forest to address the multicollinearity concern.

The R-squared values for these methods range from 0.74 to 0.78. The Gunning Fog Index appears to be the most significant predictor of passage grades as determined by subject matter experts (SMEs). FKGL follows in importance, although its regression coefficients may vary in direction depending on the originating tool. Lexile-based grade consistently rank as important across all methods. These findings indicate that complexity indices, reflecting average sentence and word lengths, complex words, and word frequency, account for up to 78% of the variance in SME-rated passage grades.

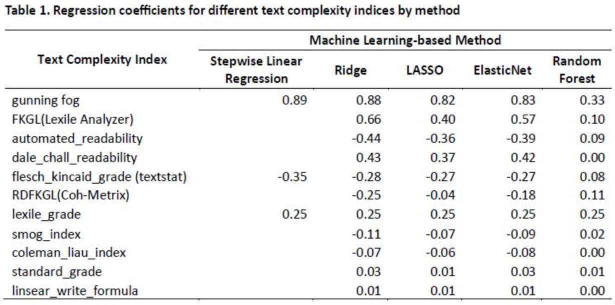


Table 1. Regression coefficients for different text complexity indices by method

Text Complexity Index	Machine Learning-based Method				
	Stepwise Linear Regression	Ridge	LASSO	ElasticNet	Random Forest
gunning fog	0.89	0.88	0.82	0.83	0.33
FKGL(Lexile Analyzer)		0.66	0.40	0.57	0.10
automated_readability		-0.44	-0.36	-0.39	0.09
dale_chall_readability		0.43	0.37	0.42	0.00
flesch_kincaid_grade (textstat)	-0.35	-0.28	-0.27	-0.27	0.08
RDFKGL(Coh-Metrix)		-0.25	-0.04	-0.18	0.11
lexile_grade	0.25	0.25	0.25	0.25	0.25
smog_index		-0.11	-0.07	-0.09	0.02
coleman_liau_index		-0.07	-0.06	-0.08	0.00
standard_grade		0.03	0.01	0.03	0.01
linsear_write_formula		0.01	0.01	0.01	0.00

Regression coefficients.jpg

Unreliability and historical bias and their relationship to test fairness and equity

Friday, 19th July - 11:45: AI and Machine Learning in Educational Settings (NB C) - Oral

Prof. Alison Cheng (University of Notre Dame), Dr. Cheng Liu (University of Notre Dame)

In this presentation we will show with both analytical derivation and simulations that the observed predictive bias of a test and test inequity in selection ratios between subgroups of the test taker population may NOT be an indicator of bias caused by or inherent in the test itself. Instead, they are artifacts of both historical bias (which may have nothing to do with the test itself) and unreliability of a test (which can be improved but never eliminated). In connecting the results to the widely used Taylor-Russell table and fairness in machine learning literature, we point out strong implications for the use and interpretation of test scores in high-stakes contexts, such as college admissions and personnel selection, especially in a currently hostile environment against testing.

Exploring the frontier of fairness: The significance of outliers in algorithmic bias and equity

Friday, 19th July - 12:00: AI and Machine Learning in Educational Settings (NB C) - Oral

Dr. Secil Ugurlu (Hacettepe University)

Decision-making, crucial for predictions in fields like educational sciences and psychology, increasingly relies on machine learning (ML) and artificial intelligence (AI), raising concerns about fairness and algorithmic bias—an unfair advantage to certain subgroups by automated decision-making algorithms. Outliers can exacerbate this bias, affecting the decisions' reliability and fairness. Thus, understanding the impact of outliers on algorithmic bias is key to ensuring decision-making processes are accurate and fair. Some AI models, such as random forests, are considered robust against various factors, including outliers. The advent of Differential Algorithmic Functioning (DAF) methods marks a significant advance, providing tools to assess and address bias in algorithmic decisions.

This study aims to investigate the effect of outliers on algorithmic bias across three distinct artificial intelligence models: Random Forests, Neural Network, and Decision Tree, using a real dataset from a higher education institution to predict decisions about academic performance of students. By comparing the outcomes of these models with and without the inclusion of outliers, this research endeavors to highlight the potential impact of outliers on the fairness and accuracy of algorithmic decisions. Utilizing Mantel-Haenszel-DAF (MH-DAF), Logistic Regression-DAF (LR-DAF), and Residual-based-DAF (R-DAF) methods (Suk & Han, 2023), the study provides a comprehensive analysis of how different approaches to handling outliers can influence the detection and mitigation of algorithmic bias. This research contributes to the ongoing discourse on ensuring fairness in algorithmic decision-making by offering empirical insights into the robustness of various AI models against outliers and their role in exacerbating or mitigating algorithmic bias.

Methodological Advances in Bayesian Graphical Modeling

Friday, 19th July - 11:30: Symposium: Methodological Advances in Bayesian Graphical Modeling (NB D) - Symposium Overview

Ms. Sara Keetelaar (University of Amsterdam)

Network psychometrics uses undirected graphical models to model the network structure of complex psychosocial systems. In recent years, the Bayesian approach to the analysis of psychological networks has gained ground as it allows the uncertainty in psychometric networks to be properly quantified. Estimating the causal structure of a psychosocial system from correlational data is extremely difficult, so the field has focused instead on estimating the conditional independence and dependence structure. In this context, Markov Random Field (MRF) models are an important class of undirected graphical models because their parameters provide direct information about the conditional independence structure of the underlying system. This symposium presents a series of methodological contributions to the Bayesian analysis of psychometric network analysis. In particular, we will introduce Bayesian graphical modeling and discuss three Bayes factor tests for conditional independence (talk 1), present results from a large-scale Bayesian re-analysis of previously published networks (talk 2), introduce a test to determine the density of a network structure (talk 3), and present a novel approach to dynamic modeling of idiographic networks (talk 4). Together, these contributions offer a comprehensive overview of current methodological challenges and innovations in Bayesian graphical modeling, aiming to improve the accuracy and interpretability of network models in psychology.

Testing conditional independence in psychometric networks

Friday, 19th July - 11:30: Symposium: Methodological Advances in Bayesian Graphical Modeling (NB D) - Symposia

Nikola Sekulovski (University of Amsterdam)

Network psychometrics uses graphical models to assess the network structure of psychological variables. An important task in their analysis is determining which variables are unrelated in the network, i.e., are independent given the rest of the network variables.

Thus, it is crucial to have an appropriate method for evaluating conditional independence and dependence hypotheses. Bayesian approaches to testing such hypotheses allow researchers to differentiate between absence of evidence and evidence of absence of connections (edges) between pairs of variables in a network. Three Bayesian approaches to assessing conditional independence have been proposed in the network psychometrics literature, however, their theoretical foundations are not widely known. In this presentation we aim to provide a conceptual review of these three methods and highlight their strengths and limitations through presenting results from a simulation study.

How robust are psychometric networks?

Friday, 19th July - 11:30: Symposium: Methodological Advances in Bayesian Graphical Modeling (NB D) - Symposia

Ms. Karoline Huth (University of Amsterdam)

Network analysis has become extremely popular in all psychological disciplines, especially in clinical psychology. Along with the enthusiasm for networks, there has been growing concern about the stability of their results. In interpreting results and accumulating our understanding across studies, we need to be aware of the potential uncertainty underlying the estimated networks. In this project, we aimed to assess the robustness of published psychological networks. The project involved a large-scale search, acquisition of datasets, and re-analysis of published cross-sectional networks. To date, we have screened 4,700 articles, invited the contribution of 1,430 papers, received 450 responses, downloaded 140 datasets, and cleaned 200 networks. We re-analyzed the collected data to address questions such as: Is there evidence of conditional independence when an edge is missing from an estimated network, or is the edge simply too unstable to be included? To adequately assess the robustness of the networks, we used a Bayesian approach because it allows us to quantify the uncertainty of the networks. In this talk, we present the re-analysis project and provide some initial insights into the robustness of the field of psychometric networks.

Are Psychological Networks Sparse or Dense?

Friday, 19th July - 11:30: Symposium: Methodological Advances in Bayesian Graphical Modeling (NB D) - Symposia

Ms. Sara Keetelaar (University of Amsterdam)

Are psychological networks sparse or dense? The answer to this question is pivotal to network psychometrics as according to network literature, sparsity is related to vulnerability of a psychological system. There is no clear consensus on what exactly is sparse, and what is dense. A review of many network results published in psychology shows that many edges are presumed to be absent. However, it is difficult to conclude if these networks are actually sparse, since results are often confounded by regularization-based estimation methods, which are based on the idea that there is only a subset of available edges present. If a sparse network is assumed, it is not surprising that we find a sparse network.

This talk will introduce a method that makes it actually possible to test if a network is sparse or dense. The test is based on Bayesian hypothesis testing, where we adopt two hypotheses: that the network is sparse and that the method is dense. Under both hypotheses, the network structure is estimated, which results in a Bayes Factor that gives evidence for either one of the hypotheses. The method can be used for different types of Markov Random Fields, depending on the type of variables. The test is implemented in the R package easybgm. In the talk, the novel method will be introduced and explained, and results will be shown.

Dynamic conditional networks for intensive repeated data

Friday, 19th July - 11:30: Symposium: Methodological Advances in Bayesian Graphical Modeling (NB D) - Symposia

Prof. Philippe Rast (University of California, Davis), Prof. Mijke Rhemtulla (University of California, Davis)

Current multilevel network analyses for intensive longitudinal data often rely on a VAR framework, leading to two types of networks: A contemporaneous, and a temporal network. Thereby, the contemporaneous network is considered constant, implying a time-invariant partial correlation network throughout all time points. We introduce a novel multilevel network modelling strategy that deviates from the notion of a static contemporaneous network. Our proposed method constructs dynamic conditional contemporaneous networks for each time point, accommodating potential variations in partial correlations across time and individuals. Specifically, our approach merges a multilevel VAR structure for the mean with a multilevel dynamic process for the scale, allowing the residuals to be heteroskedastic. This implies that the residual covariance structure, and with it the contemporaneous network, at time t is modelled conditionally on the preceding contemporaneous network at time t-1. Similar to the VAR portion of the model, the dynamic conditional correlation parameters for the contemporaneous network are defined by fixed and random effects. The model is illustrated using the iFit study, an ecological momentary assessment (EMA) on daily health behaviours and physical health outcomes over 100 days.

Applying Fisher’s method of scoring to estimate ability parameters of the multi-unidimensional pairwise-preference model

Friday, 19th July - 11:30: Estimation Methods (RB 209) - Oral

Dr. Yeh-Tai Chou (Research Center for Educational and Psychological Testing, National Taiwan Normal University), Prof. Ching-Lin Shih (Institute of Education & Center for Teacher Education, National Sun Yat-sen University), Prof. Yao-Ting Sung (Department of Educational Psychology and Counseling, National Taiwan Normal University)

Socially desirable responding (SDR) is commonly observed in self-report measures, posing challenges to test validity. In response to this challenge, the force-choice (FC) format, such as pairwise-comparison items, was developed to mitigate SDR behavior. The multi-unidimensional pairwise-preference (MUPP) model (Stark, Chernyshenko, & Drasgow, 2005) was proposed to analyze responses to pairwise-comparison items. However, due to the model’s complexity, the ability parameters were estimated using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method with approximated inverse Hessian matrices. This approach can lead to issues such as higher estimated standard errors of measurement and convergence difficulties during parameter estimation. These issues restrict the applicability of the MUPP, particularly in computerized adaptive testing contexts. To address these issues, this study applies the Newton-Raphson method, integrating Fisher’s scoring procedure by replacing Hessian matrices with the Fisher information matrix, to estimate ability parameters of the MUPP model. The method’s performance in terms of parameter recovery and convergence percentage was evaluated through simulation studies. Results demonstrate that the Newton-Raphson method efficiently achieves convergence and provides reasonable standard errors of measurement, potentially enhancing the applicability of the MUPP model in computerized adaptive testing. This study discusses the implications of its findings and explores potential future applications in psychometrics.

Keywords: Force-choice, MUPP model, Newton-Raphson, Fisher’s method of scoring

Estimating hierarchical drift diffusion models: a comparison of methods

Friday, 19th July - 11:45: Estimation Methods (RB 209) - Oral

Ms. Yufei Wu (KU Leuven), Mr. Kristof Meers (KU Leuven), Prof. Francis Tuerlinckx (KU Leuven)

The drift diffusion model (DDM) is a prominent model in cognitive science and decision-making research (Ratcliff, 1978; Ratcliff et al., 2016). Its hierarchical extension allows researchers to study individual differences in the psychological processes underlying decision making (Johnson et al., 2017; Vandekerckhove et al., 2011). Nonetheless, the inherent complexity of the DDM combined with the hierarchical structure make the estimation of its parameters challenging. Recently, various methods and associated software packages have been developed to fit this model, employing Bayesian frameworks such as Markov Chain Monte Carlo sampling (i.e., in JAGS; Wabersich & Vandekerckhove, 2014) and deep learning algorithms (i.e., in BayesFlow; Radev et al., 2022). However, a comprehensive comparison of these packages is lacking. In this study, we propose an in-depth comparison of the methods and associated software packages using simulations , and an evaluation of each method and package in terms of its flexibility, computation speed, and parameter recovery. By comparing these aspects of different software packages, the study aims to provide insights into their relative strengths and weaknesses, helping researchers choose the most appropriate software for their specific needs when working with the hierarchical DDM.

Dealing with small-sample biases in AR(1) models: Methods for psychological research

Friday, 19th July - 12:00: Estimation Methods (RB 209) - Oral

Zhiwei Dou (KU Leuven), Dr. Sigert Ariens (KU Leuven), Prof. Eva Ceulemans (KU Leuven), Prof. Ginette Lafit (KU Leuven)

Intensive longitudinal studies are state-of-the-art designs to study the dynamics of daily life psychological processes. These designs yield time series of data, which are commonly analyzed using first-order autoregressive [AR(1)] modeling.In the AR(1) model, a variable is a function of its own value at the previous time point (i.e., autoregressive effect). the Ordinary Least Squares (OLS) estimator is a common application for estimating the autoregressive effect. It is asymptotically consistent, but it is downward biased when the sample size is small, which is common in intensive longitudinal designs. In this talk, we will showcase using simulation studies the performance of alternative methods that might improve the estimation of the autoregressive effect when the sample size is limited. Specifically, we will focus on two types of approaches. In the first approach, the OLS estimator is corrected based on the analytical correction formula proposed by Kendall (Marriott and Pope, 1954 and Kendall, 1954). In the second approach, a Bayesian correction with a prior about initial observables (Jarocinski and Marcet, 2010) is used to estimate the AR(1) model autocorrelation effect. These approaches will also be compared with other estimation methods which are often employed in psychological applications, such as the popular Bayesian single subject AR(1) model (Schuurman et al., 2015). Finally, the proposed methods will be applied to real intensive longitudinal data sets to the extent to which the different methods impact empirical findings.

Empirical bias-reducing adjustments for Item Response Theory (IRT) models

Friday, 19th July - 12:15: Estimation Methods (RB 209) - Oral

Dr. Haziq Jamil (Universiti Brunei Darussalam), Prof. Ioannis Kosmidis (University of Warwick)

In the field of psychometrics, the accuracy and reliability of measurement tools are paramount, particularly when employing Item Response Theory (IRT) models for assessing latent psychological traits. A persistent challenge in this domain is the non-zero bias of order O(1/n) in finite sample sizes, a problem aggravated by deviations from the latent normality assumption, such as excess zeroes or skewed distributions. This presentation introduces an empirical bias adjustment method designed to mitigate this problem. The method applies adjustments derived from the empirical approximation of bias through higher-order derivatives of the estimating functions. Our simple approach offers a promising avenue for enhancing the robustness of IRT model estimations, especially in samples that deviate from idealized assumptions. The method's theoretical advantages include markedly improved accuracy of estimator recovery, rendering it an invaluable asset for both researchers and practitioners. The innovation lies in its straightforward adjustment process, which can be implemented via implicit (i.e. solving adjusted estimating equations) or explicit methods (i.e. adjusting original estimators), thus streamlining the adoption and offering an appealing alternative to existing, more complex bias-reduction techniques. Validation of our theoretical framework through simulation studies confirms the effectiveness of our empirical bias adjustment in reducing parameter bias, thereby enabling more precise and dependable psychometric measurements.

Estimating high-dimensional latent variable models via minibatch variance-reduced stochastic optimisation

Friday, 19th July - 12:30: Estimation Methods (RB 209) - Oral

Mr. Motonori Oka (The London School of Economics and Political Science), Dr. Yunxiao Chen (The London School of Economics and Political Science), Prof. Irini Moustaki (The London School of Economics and Political Science)

Latent variable models are widely for analysing data in social and behavioural sciences including education, psychology, and political science. In recent years, high-dimensional latent variable models have been used extensively to analyse large and complex data. However, the marginal maximum likelihood estimation of high-dimensional latent variables poses significant computational bottlenecks due to the high complexity of integrals of latent variables. To tackle this issue, stochastic optimisation, a method combining stochastic approximation and sampling techniques, has been proven as a powerful way to address this computational challenge. This method involves two steps: first, sampling the latent variables from their posterior distribution given the current parameter estimate, and second, updating the fixed parameters using an approximate stochastic gradient constructed by plugging in the sampled latent variable samples. In our research, we propose a computationally more efficient stochastic optimisation algorithm. Especially, the proposed algorithm achieves improvement through three main elements: employing a minibatch of observations when sampling latent variables and constructing stochastic gradients, utilising an unadjusted Langevin sampler for sampling latent variables, and implementing a variance reduction trick for constructing stochastic gradients. We also establish the convergence of our proposed algorithm under suitable regularity conditions. Simulation studies involving confirmatory 2-parameter logistic item response and multilevel logistic regression models with random effects show that our proposed algorithm performs better than several competitors. The proposed method is further applied to a real-world dataset in personality assessments for empirical illustration.

Item compromise and preknowledge detection using response time and distractors

Friday, 19th July - 11:30: Topics in IRT 4 (RB 210) - Oral

Dr. Merve Sarac (College Board)

Research on detecting item compromise and examinee preknowledge has recently incorporated distractor selection into the detection methods in multiple-choice tests. However, statistical evidence from response times, available in computer-based assessments, remains underutilized for identifying compromised test content. This study uses item fit analysis based on response times alongside item scores and distractor selection to detect potentially compromised items. Item fit analyses are based on a residual statistic, a χ^2 statistic, and a test of normality suggested for item scores, distractors, and response times, respectively. Leveraging information from misfitting items, a detection statistic based on item scores and response times – the signed likelihood ratio test – is used to detect examinee preknowledge. In simulations, using several design factors (e.g., the number of compromised items, the number of examinees with preknowledge, and disclosed key accuracy), the false positive rate and true positive rate are evaluated for both items and examinees. Preliminary results indicate that incorporating response time in item fit analysis alongside item scores and distractor selection improves the true positive rates for both items and examinees.

Exploratory higher-ordered diagnostic modeling framework for general response types

Friday, 19th July - 11:45: Topics in IRT 4 (RB 210) - Oral

Ms. Jia Liu (Northeast Normal University), Dr. Yuqi Gu (Columbia University)

Cognitive Diagnostic Models (CDMs) are essential tools in educational assessment, offering a structured approach to diagnosing examinee proficiency based on their responses to test items. While current CDMs mainly focus on binary or categorical responses, there's a growing need to extend their use to encompass a wider range of response types, including continuous and count-valued responses. Additionally, integrating higher-order latent trait structures has become crucial for gaining deeper insights into cognitive concepts and traits. In this paper, we propose a modeling framework for general-response higher-ordered CDMs. Our framework features a highly flexible bottom layer capable of adapting to various response types and a variety of measurement models for CDMs. We address a more challenging exploratory analysis scenario where the item-attribute relationship, conventionally specified by the Q-matrix, is unknown and needs to be estimated along with the other model parameters. In the higher-order layers, we employ a probit-link based model, highlighting its benefits in terms of identifiability and computational efficiency. We present identifiability results, serving as foundational guidelines for optimizing test design and enhancing the accuracy of our model. For estimation and computation, we propose two efficient EM-type algorithms based on different sampling strategies and conduct simulation studies to demonstrate their efficacy.

Using the multidimensional nominal response model to model faking: The importance of item desirability characteristics

Friday, 19th July - 12:00: Topics in IRT 4 (RB 210) - Oral

Mr. Timo Seitz (University of Mannheim), Prof. Eunike Wetzel (RPTU Kaiserslautern-Landau), Prof. Benjamin E. Hilbig (RPTU Kaiserslautern-Landau), Prof. Thorsten Meiser (University of Mannheim)

When self-report personality questionnaires are used in high-stakes assessments, there is the risk that test-takers present themselves in an overly favorable manner, that is, engage in faking. Unless the influence of faking on item responses is taken into account, faking can harm the psychometric properties of a test. In the present research, we account for the influence of faking using an extension of the multidimensional nominal response model (MNRM) with item-specific scoring weight vectors. Particularly, we investigated under which conditions the MNRM can adequately adjust substantive trait estimates and other model parameters for faking. Because of potential collinearity between scoring weight vectors of faking and substantive traits, we hereby focused on the role of variation in the way item content is related to social desirability (i.e., item desirability characteristics) in facilitating the modeling of faking and counteracting its detrimental effects. Using a simulation, we found that the inclusion of a faking dimension in the model can overall improve the recovery of substantive trait person parameters and other model parameters, especially when the impact of faking in the data is high. Item desirability characteristics moderated the effect of modeling faking and were themselves associated with different levels of parameter recovery. In an empirical demonstration, we also showed that the faking modeling approach in combination with different item desirability characteristics can prove successful in empirical questionnaire data. These findings have important implications for the psychometric modeling of faking in different contexts.

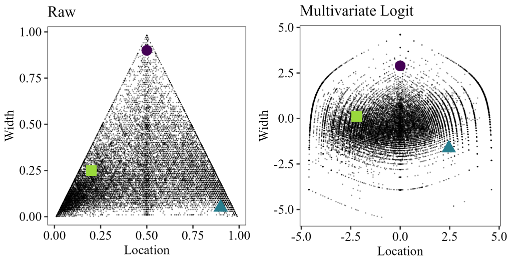
A multivariate logit-function for modeling continuous bounded interval responses

Friday, 19th July - 12:15: Topics in IRT 4 (RB 210) - Oral

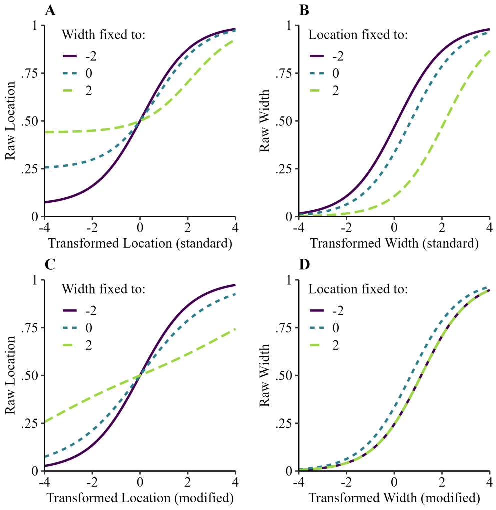
Mr. Matthias Kloft (Psychological Methods Lab, Department of Psychology, University of Marburg), Prof. Daniel W. Heck (Psychological Methods Lab, Department of Psychology, University of Marburg)

Interval response formats allow respondents to indicate a range of values by setting a lower and an upper bound. This can be used, for example, to quantify the uncertainty of an estimate of a particular quantity or to assess the variability of stimuli. However, analyzing continuous bounded interval responses presents statistical challenges due to stochastic dependencies inherent in this response format: First, the two interval bounds are constrained by the lower and upper scale limits, respectively. Second, the two interval bounds are mutually constrained, with the lower bound necessarily being smaller than the upper bound. These dependencies can lead to skewed distributions of marginal responses. Also, to provide interval responses close to the scale limits, a respondent has to reduce the width of the response interval, resulting in a nonlinear negative relationship between interval location and width.

For continuous bounded responses, the logit function is often used to transform responses into a more appropriate form for linear models that assume a normal distribution. However, the standard multivariate version of the logit function does not respect the symmetry property of interval responses and therefore lacks the desired interpretation of the transformed values in terms of location and width. We propose a modified multivariate logit function that maintains this intuitive interpretation. We illustrate the application of this novel approach to empirical examples and show that it achieves better model fit compared to the modeling of untransformed interval locations and widths.



Logit vs raw.png



Inv logit trace.png

Respondent-level change-point detection in educational and psychological testing

Friday, 19th July - 12:30: Topics in IRT 4 (RB 210) - Oral

Dr. Gabriel Wallin (Lancaster University), Dr. Yunxiao Chen (The London School of Economics and Political Science), Prof. Xiaoou Li (University of Minnesota, Twin-Cities)

Modern educational and psychological testing data are increasingly heterogeneous. This study focuses on individual-level changes in response style that happen during the test due to, for example, time pressure in high-stakes testing or lack of motivation in low-stakes testing. To better model the underlying response process, we introduce a novel statistical methodology where a baseline Item Response Theory (IRT) model is fused with a change-point model to characterize such respondent-level changes. The baseline IRT model (e.g. the 2PL model) captures normal item response behaviour, and the post-change model captures the aberrant behaviours (e.g., item responses under time pressure). Every respondent has either none or one change, but the proposed framework can be generalized to accommodate multiple change-points. The change-points are treated as latent variables whose change-point probability depends on the latent construct level of the respondent and item parameters. The joint model can thus be inferred under the standard statistical framework for latent variable models. We propose an Expectation-Maximization algorithm that estimates the baseline and change-point model parameters simultaneously. The proposed method is evaluated in a comprehensive simulation study, which demonstrates its effectiveness in detecting the respondents with a change-point and the location of the change-point, whilst simultaneously estimating the rest of the model parameter. We also showcase the method in an application to real-world ability testing data.

Researching response-scale format effects in questionnaires: Using the Height Inventory

Friday, 19th July - 11:30: Psychometric Applications to Health Outcomes 2 (RB 211) - Oral

Dr. Hynek Cígler (Masaryk University), Dr. Stanislav Ježek (Masaryk University), Mr. Karel Rečka (Masaryk University)

Self-report measures of attitudes and personality characteristics mainly use items with Likert-type response scales (LS) where respondents select an answer from a range of ordered options (e.g., agree–disagree). Such response scales can differ in several formal attributes – number of options, presence of verbal anchors, their extremity, or orientation (so-called reversed items). These may affect the reliability and validity of responses and total scores (Furr, 2011).

To assess response-format effects, our team extensively exploits the unique properties of our Height Inventory (Rečka, 2018). The idea of measuring height using a psychological questionnaire is not entirely original (van der Linden, 2016; Kam et al., 2021). However, we elaborated a new methodological approach utilizing self-reported height to assess the criterion validity of observed scores and latent traits using latent variable models and measurement invariance analysis.

So far, we used this approach to study the effects of LS length, the presence and extremity of its verbal anchors, and reversed-key items. We also assessed the performance of the Visual Analogue Scale compared to LS and the effects of speedy vs. careful responding. Some of our findings contradict common beliefs; for example, that binary items might have higher criterion validity than longer response scales.

Our talk briefly outlines the Height Inventory and its psychometric properties. Then, we describe our methodological approach in detail, focusing on its advantages and limits.

Bot detection: Simulations and application in people-centered health measurement surveys

Friday, 19th July - 11:45: Psychometric Applications to Health Outcomes 2 (RB 211) - Oral

Dr. Carl Falk (McGill University), Mr. Michael John Ilagan (McGill University), Amaris Huang (McGill University), Mathilde Verdam (Leiden University), Richard Sawatzky (Trinity Western University)

In this research, we present recent work on an algorithm developed by Ilagan & Falk (in press) to detect random responding (e.g., by survey bots) to Likert-type items. The algorithm requires neither a measurement model nor an a priori sample of known humans and bots. While the algorithm maintains a nominal 95% sensitivity for detecting bots (based on a permutation test), its classification accuracy may depend on inventory- or sample-specific factors (e.g., modes of data collection). To understand conditions that affect the algorithm's classification accuracy, we briefly review recent simulation-based research suggesting that its performance is better with more item responses, more categories per item, and more variability in the IRT difficulty parameters of the survey items. Inspired by a large dataset measuring pain and emotional well-being for a study on advancing equitable people-centered health measurement, we then tackle the problem of missing data due to non-response or "don't know" or "prefer not to answer" responses. We present simulations to understand the algorithm's performance under conditions designed to mimic this application. In brief, classification accuracy was higher with more complete item responses. Challenges in detecting bots and random responders in the context of this application are also discussed, such as computational issues with a large dataset, heterogeneity in the response process for humans, whether bots may create missing data, and trade-offs between sensitivity and specificity.

Cognitive psychometric approach for modeling cognitive aging across ambulatory assessments

Friday, 19th July - 12:00: Psychometric Applications to Health Outcomes 2 (RB 211) - Oral

Sharon Kim (The Pennsylvania State University), Lindy Williams (The Pennsylvania State University), Dr. Michael Hunter (The Pennsylvania State University), Dr. Zita Oravecz (The Pennsylvania State University)

Technological innovations (e.g., smartphone applications) make it possible to collect high-frequency repeated assessments about people's naturalistic, real-time cognitive functioning. Fitting cognitive psychometric models to such intensive longitudinal data can disentangle the underlying dynamic processes generating manifest cognitive performance. In this talk, we introduce a Bayesian exponential learning process model, which captures individual differences in cognitive performance in terms of four parameters: learning rate, retest gain, peak performance, and performance inconsistency. We cast the model in a multilevel framework which allows for exploring sources of individual differences by regressing these four parameters on predictors in a single-step analysis. We will demonstrate the utility of this model by analyzing data from the Einstein Aging Study. Results showed that participants with mild cognitive impairment had credibly lower peak performance, higher retest gain and more performance inconsistency across two cognitive domains. These are promising results towards using the parameters of the proposed model as novel digital markers for risk of cognitive decline.

Do psychometrics matter? The effects of psychometrics on depression trial outcomes

Friday, 19th July - 12:15: Psychometric Applications to Health Outcomes 2 (RB 211) - Oral

Dr. David Byrne (Royal College of Surgeons Ireland), Prof. Frank Doyle (Royal College of Surgeons Ireland), Dr. Fiona Boland (Royal College of Surgeons Ireland)

Objective

It has been contended that the sophisticated statistical techniques that are used to evaluate psychometric scales are vital to improving psychometric assessment. However, the implications of these techniques may not be fully appreciated by applied researchers. Additionally, they can often provide conflicting results and there is limited evidence that adopting these techniques actually makes important differences to ultimate outcomes. We therefore aimed to determine whether applying psychometric analyses to individual patient data would demonstrate important differences in depression trial treatment effects.

Methodology

We conducted a secondary analysis of individual participant data from 15 antidepressant treatment trials from Vivli.org (n=6962) that used the Montgomery-Asberg Depression Rating Scale (MADRS). Pooled data was analysed using confirmatory factor analysis, item response theory and network analysis, providing psychometrically-informed model scores to compare to original total MADRS scores in multilevel models. Differences in trial effect sizes was the outcome of interest.

Results

The MADRS performed well under psychometric evaluation post-treatment, with some issues noted for pre-treatment models. Optimal models also differed across psychometric methods. Effect size analyses of psychometrically informed models showed no difference in outcome when compared to original trial effect sizes.

Conclusion

This is the first large, multi-trial application of psychometric analyses to randomised trials. The methods did not moderate treatment effects, which may be due to the small effects inherent in these trials. If replicated, results suggest that the number of items in measurement scales could be reduced without affecting the power to detect treatment effects, which may reduce patient burden.

#LoveIrishResearch

Probabilistic Forecasting with International Large-Scale Assessments: Methods for Estimating the Pace of Progress to the UN Education Sustainable Development Targets

Friday, 19th July - 14:00: Presidential Address (Vencovského aula) - Presidental Address

Prof. David Kaplan (University of Wisconsin - Madison)

In 2015, the United Nations adopted the Sustainable Development Goals (SDGs). Regarding education, the UN identified equitable, high-quality education, including the achievement of literacy and numeracy for all youth and adults as a key goal. To assess country-level progress toward these goals, it is important to monitor trends in educational outcomes over time. This talk demonstrates how optimally predictive growth models can be constructed in order to monitor the pace of progress at which countries are progressing toward the education SDGs. A number of growth curve models can be specified to estimate the pace of progress, however, choosing one model and using it for predictive purposes assumes that the chosen model is the one that generated the data. A classical approach to addressing this type of model uncertainty is Bayesian model averaging (BMA). However, BMA rests on the assumption that the true data generating model is in the set of models that are being averaged. In this talk, we adapt and apply the ensemble prediction method of Bayesian stacking to form mixtures of predictive distributions from an ensemble of individual models specified to predict country-level growth rates. Bayesian stacking relaxes the true data generating model assumption. We demonstrate Bayesian stacking on country-level data from PISA. Our results show that Bayesian stacking yields better predictive accuracy than any single model as measured by the Kullback-Leibler divergence. On the basis of the ensemble average calculated from the stacked predictive distribution, we show a forecast plot for one PISA cycle ahead.

IMPS 2024

Prague, Czech Republic

July 16-19, 2024 • Short Courses July 15

ABSTRACT BOOK: POSTERS

Estimating the Root Mean Square Error of Approximation (RMSEA) with Multiply Imputed Data under Nonnormality

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Yunhang Yin (University of South Carolina), Prof. Dexin Shi (University of South Carolina), Prof. Amanda Fairchild (University of South Carolina)

Multiple imputation (MI) is one of the recommended modern techniques for handling missing data in structural equation modeling (SEM), and evaluating model fit is a crucial aspect of analyzing SEM models. Methods for pooling model fit indices across imputed datasets are still under development, however, especially in the context of nonnormal data. In this simulation study, we considered methods for estimating a robust measure of the Root Mean Square Error of Approximation (RMSEA) fit index, and introduced strategies for pooling the robust RMSEA across imputed datasets. We evaluated the performance of these strategies under various conditions by manipulating sample size, level of nonnormality, nonnormal data generation algorithm, missing data mechanism, and percentage of missing data. Results showed that a MI-based RMSEA built upon Lai (2020) tended to outperform other approaches, yielding smaller bias in point estimates. Furthermore, by using a normal approximation, confidence intervals (CIs) for the population RMSEA can be computed with better coverage rates. Drawing on our findings, we discuss the practical implications of our study and suggest directions for future research.

Bayesian Pattern-mixture Model for Nonignorable Missing Data in Longitudinal Mediation Analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Suyu Liu (University of Texas MD Anderson Cancer Center)

Missing data are common in longitudinal studies. We propose a Bayesian pattern-mixture model to handle nonignorable or informative missing data for longitudinal mediation analysis. We partition the observed longitudinal data into different missing patterns. Within each pattern, the Bayesian hierarchical model is used to model the mediation analysis. As each of the variables in the mediation structure may have different missing patterns, this may result in a large number of missing patterns. We proposed a latent-class model to collapse the different missing-data patterns for parsimonious inference. The simulation study shows that the proposed method has desirable operating characteristics and effectively eliminates the bias due to nonignorable missing data.

A Two-step estimator for growth mixture models with covariates

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Yuqi Liu (Leiden University), Dr. Zsuzsa Bakk (Leiden University), Dr. Ethan McCormick (Leiden University), Prof. Mark de Rooij (Leiden University)

Growth mixture models (GMMs) are a popular approach for modeling unobserved population heterogeneity in change over time by assuming that the change trajectories are described by a set of latent classes (LCs). GMMs can be extended to incorporate covariates, which can predict LC membership, the within-class growth trajectories, or both. However current estimators are sensitive to misspecifications in complex models. We propose extending the two-step estimator for LC models (Bakk & Kuha, 2018) to GMMs, which provides robust estimation against model misspecifications for simpler LC models. We conducted a series of simulation studies, comparing the performance of the newly proposed two-step estimator to the commonly-used one- and three-step estimators (Diallo & Lu, 2017). Two different simulation conditions were considered, 1) the population model only included covariates that predicted the LC membership, and 2) the covariates predicted both LC membership and the within-class growth factors. Additionally, we varied strength of the measurement model and number of time points.

Results show that when predicting LC membership alone, all three estimators are unbiased when the measurement model is strong, with weak measurement model results being more nuanced. Alternatively, when including covariates effects on the measurement model growth factors, the two-step approach performs better than the three-step estimator, and comparably to the one-step estimator. The two-step estimator shows consistent robustness against misspecifications across simulation conditions. The three-step estimator performs worth with weak measurement models than the other approaches, while the one-step approach is most sensitive to misspecifications

Culturally responsive testing using generative AI

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Prof. Youn-Jeng Choi (Ewha womans university), Ms. Jiyoon Lee (Ewha womans university), Dr. Jaehwa Choi (George Washington University)

It has been considered that a good question could be solved entirely by the student's abilities without being affected by the student's sociocultural background. Since producing good quality items takes a lot of time and money, test developers have tried to make items as fair and unbiased as possible. However, it is not easy to develop unbiased and fair questions that are not affected by all races, cultures, environments, etc., and with the advent of generative AI, it is possible to make items that respond culturally (Choi, 2023). This study aims to illustrate how to create culturally responsive items for immigrant students and to evaluate them using ChatGPT. We will make items based on common core standards from the USA and develop culturally responsive parallel items using ChatGPT4 for students from Vietnam, China, and South Korea. We will perform the simulation studies using various test contents, item format, test length, and so on with 20 replications to evaluate the quality/accuracy of item generation, test fairness, and bias against cultures. The evaluation of the test items so far has mainly consisted of analyzing student responses based on measurement theory by psychometricians and content validity verification by content experts. Since students' responses are required for item analysis, it is difficult to evaluate the test before students attempt the test. With the help of ChatGPT, we can evaluate items simultaneously with item development. Through this study, we will find out how consistently and accurately ChatGPT evaluates items.

Cultural and linguistic DIF in PISA 2018 reading assessment

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Yejin Woo (Ewha womans university), Ms. Ji Yoon Eom (Ewha womans university), Prof. Youn-Jeng Choi (Ewha womans university)

The Programme for International Student Assessment (PISA) is an assessment conducted among students from diverse linguistic and cultural backgrounds and inevitably faces challenges due to the impact of language and culture on testing outcomes (Huan, Wilson, & Wang, 2016). Our research investigates how these factors influence Differential Item Functioning (DIF) in reading assessments, examining bias at both the test and item levels.

This study employs a comparative analysis with the United States as a reference, contrasted with Canada, Singapore, and South Korea, focusing on responses to the PISA 2018 'Rapa Nui' unit. We balanced sample sizes across nations (n=2,816) to mitigate model fit sensitivity to sample size fluctuations (Lei & Lomax, 2005; Fan & Sivo, 2007; Mahler, 2011). After preprocessing data, recoding partial credit responses, and addressing missing values, DIF was analyzed using IRT-likelihood Ratio Test (IRT-LR), Rasch-tree model, and Logistic regression model with IRTLRDIF and R computer programs.

The results from IRT-LR and logistic regression indicate the strongest DIF effects in Korea, Singapore, and Canada, in that order, compared to the United States, confirming the significant impact of language and culture on responses (see Table 1). Rasch-tree analysis further identifies specific cultural factors, such as competitiveness and the perception of teacher's stimulation of reading engagement, as significant to DIF (see Figure 1). This implies the necessity for precise adjustments for linguistic and cultural differences to ensure measurement invariance. Future analysis will further examine the actual linguistic and cultural factors influencing DIF items in the 'Rapa Nui' unit.

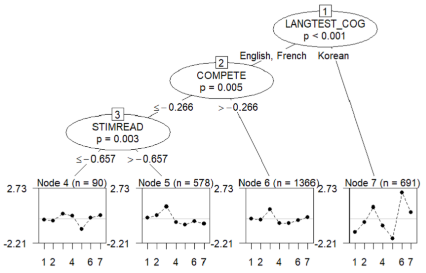


Figure 1. Rasch Tree Analysis of the Rapa Nui Unit Reading Skills Items from PISA 2018: Impact of Language of Measurement, Achievement Goals, and Reading Instruction

Figure 1.png

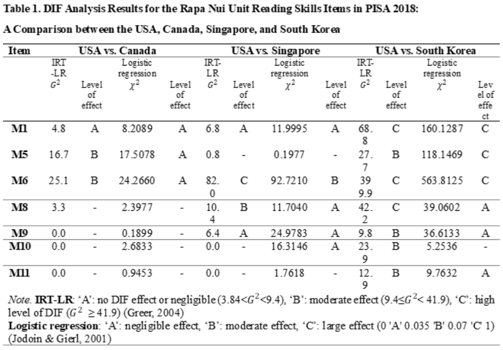


Table 1. DIF Analysis Results for the Rapa Nui Unit Reading Skills Items in PISA 2018:

A Comparison between the USA, Canada, Singapore, and South Korea

Item	**USA vs. Canada**				**USA vs. Singapore**				**USA vs. South Korea**			
	IRT-LR G^2	Level of effect	Logistic regression χ^2	Level of effect	IRT-LR G^2	Level of effect	Logistic regression χ^2	Level of effect	IRT-LR G^2	Level of effect	Logistic regression χ^2	Level of effect
M1	4.8	A	8.2089	A	6.8	A	11.9995	A	68.8	C	160.1287	C
M5	16.7	B	17.5078	A	0.8	-	0.1977	-	27.7	B	118.1469	C
M6	25.1	B	24.2660	A	82.0	C	92.7210	B	399.9	C	563.8125	C
M8	3.3	-	2.3977	-	10.4	B	11.7040	A	42.2	C	39.0602	A
M9	0.0	-	0.1899	-	6.4	A	24.9783	A	9.8	B	36.6133	A
M10	0.0	-	2.6833	-	0.0	-	16.3146	A	23.9	B	5.2536	-
M11	0.0	-	0.9453	-	0.0	-	1.7618	-	12.9	B	9.7632	A

Note. **IRT-LR**: 'A': no DIF effect or negligible (3.84<G^2<9.4), 'B': moderate effect (9.4≤G^2< 41.9), 'C': high level of DIF (G^2 ≥41.9) (Greer, 2004)
Logistic regression: 'A': negligible effect, 'B': moderate effect, 'C': large effect (0 'A' 0.035 'B' 0.07 'C' 1) (Jodoin & Gierl, 2001)

Table 1.png

Exploratory Extension of Generalized Structured Component Analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Naoto Yamashita (Kansai University)

Generalized Structured Component Analysis (GSCA) is a powerful method within Structural Equation Modeling (SEM) for constructing components from observed variables and examining their regression relationships. This research introduces an exploratory extension of GSCA, termed Exploratory GSCA (EGSCA), akin to the concept of Exploratory SEM (ESEM) in factor-based SEM procedures. EGSCA explores relationships between observed variables and components through orthogonal rotation of parameter matrices, addressing the indeterminacy inherent in GSCA. This presentation outlines the EGSCA procedure and introduces a specialized rotational algorithm tailored for EGSCA, aimed at simplifying all parameter matrices simultaneously. Through a numerical simulation studie, EGSCA's effectiveness in recovering true parameter values is demonstrated, surpassing existing GSCA procedure. Additionally, EGSCA is applied to two real datasets, revealing better model suggestions compared to previous research, thereby highlighting its efficacy in model exploration within SEM.

Longitudinal network of peer problems and emotional symptoms among adolescents

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Pedro Henrique Ribeiro Santiago (The University of Adelaide), Prof. Lisa Smithers (University of Wollongong), Dr. Michelle Townsend (University of Wollongong), Dr. Adrian Quintero (Icfes – Instituto Colombiano para la Evaluación de la Educación), Dr. Alyssa Sawyer (The University of Adelaide), Dr. Gustavo Soares (The University of Adelaide), Dr. Kym McCormick (The University of Adelaide), Ms. Alexandra Procter (The University of Adelaide), Prof. Lisa Jamieson (The University of Adelaide)

When investigating psychological networks, researchers have increasingly employed causal discovery algorithms to identify a Directed Acyclic Graph (DAG), where nodes indicate behaviours, cognitions or emotions and edges indicate causal effects (Briganti et al., 2022). However, one challenge is that several DAGs can be (more or less) compatible with the data, so Bayesian structure learning of DAGs was proposed to evaluate the posterior probability of a range of possible DAGs (Suter et al., 2021). We demonstrate the application of this method to evaluate the longitudinal network of peer problems and emotional symptoms among adolescents aged 12 to 14 years. Data was from the Longitudinal Study of Australian Children (LSAC) and samples included adolescents who participated in the Baby (n=2,694) or Kindergarten (n=3,144) Cohorts at two follow-ups (ages 12 and 14). Peer problems and emotional symptoms were measured with the Strengths and Difficulties Questionnaire. The analytical steps were: (1) Bayesian structure learning of DAGs was employed in the LSAC K Cohort and the consensus DAG (posterior edge probability > 70%) was identified; (2) since the DAG can be conceptualised as a non-parametric structural equation model (SEM) (Hernán & Robins, 2006), the consensus DAG was evaluated with Bayesian SEM (by assuming linearity for all causal effects) in an independent sample, the LSAC B Cohort; and (3) intervention targets were identified with centrality measures. The findings showed that, in the longitudinal network of peer problems and emotional symptoms among adolescents aged 12 to 14 years, reducing bullying and excessive worries constituted key intervention targets.

An improved inferential procedure to evaluate item discriminations in a conditional maximum likelihood framework

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Andreas Kurz (University of Salzburg), Dr. Clemens Draxler (UMIT TIROL Private University for Health Sciences and Technology)

In this poster session we discuss a modified and improved inductive inferential approach to evaluate item discriminations in a conditional maximum likelihood and Rasch modeling framework. A random intercept or mixed effects logit model (for binary data) is suggested which has similarities to the well-known 2PL model. Unlike the latter it constitutes a multiparameter exponential family that allows for conditional likelihood inference.

The new approach implies a linear restriction of the assumed set of probability distributions in the classical approach (i.e., an unrestricted comparison of item parameters between person groups with different scores) that represents scenarios of different item discriminations in a straightforward and efficient manner. It involves the derivation of four chi square hypothesis tests based on asymptotic theory as well as one parameter-free test suitable in small sample size scenarios. One of the tests is a modification of Andersen’s likelihood ratio test typically used to test equality of item parameters between different person score groups. We discuss the improvement of the tests and compare it to other procedures like information criteria. The results of Monte Carlo experiments as well as real data examples from educational research show an improvement of power of the modified tests of up to 0.3.

Longitudinal R-DINA Model: A Solution to Small Samples

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Hyunjee Oh (Teachers College, Columbia University), Prof. Chia-Yi Chiu (Teachers College, Columbia University)

The study aims to develop a new longitudinal cognitive diagnosis model (CDM) for measuring changes in examinees' attribute mastery growth with small longitudinal data. Based on a study by Oh and Chiu (2024), the longitudinal CDMs such as the transition diagnostic classification model (TDCM) and the latent transition analysis - deterministic input, noisy "and" gate (LTA-DINA) model do not perform better than the non-longitudinal CD methods with no longitudinal components when samples are small (i.e., $N < 100$) across various conditions. Inspired by the finding, we propose a longitudinal restricted DINA (LR-DINA) model that brings in the restricted DINA model (R-DINA; Nájera, Abad, Chiu, & Sorrel, 2023) to the longitudinal framework for small educational settings. The R-DINA model is featured in that it results in estimates of examinees' attribute patterns identical to those produced by the nonparametric classification (NPC) method, which is known to be an effective and efficient method for small samples. Hence, the inclusion of the R-DINA model takes advantage of its excellent performance with small sample and the same time, allows us to implement LTA component. To compare the performance of the proposed model with other longitudinal CDMs, systematic simulation studies with various conditions such as different test lengths, number of attributes, and item qualities will be conducted. The results from the simulations studies will be evaluated by the mean pattern-wise agreement rates (PAR) as well as CPU times. The model will be a feasible and promising alternative to existing longitudinal CDMs when samples are small.

MixML-SEM: A parsimonious approach for finding clusters of groups with equivalent structural relations in presence of measurement non-invariance

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Hongwei Zhao (KU Leuven), Prof. Jeroen Vermunt (Tilburg University), Kim De Roover (KU Leuven)

Structural equation modeling (SEM) is commonly used to explore relationships between latent variables, such as beliefs and attitudes. However, comparing structural relations across a large number of groups, such as countries, can be challenging. Existing SEM approaches may fall short, especially when measurement non-invariance is present. In this project, we propose Mixture Multilevel SEM (MixML-SEM), a novel approach to comparing relationships between latent variables across many groups. MixML-SEM gathers groups with the same structural relations in a cluster, while accounting for measurement non-invariance in a parsimonious way by means of random effects. Specifically, MixML-SEM captures measurement non-invariance using multilevel CFA and it then estimates the structural relations and mixture clustering of the groups by means of the structural-after-measurement (SAM) approach. In this way, MixML-SEM ensures that the clustering is focused on structural relations and unaffected by differences in measurement. In contrast, multilevel SEM estimates measurement and structural models simultaneously, and both with random effects. If desired, a post-hoc clustering can be applied to the group-specific regression estimates derived from the random effects. In comparison to ML-SEM, MixML-SEM proves particularly advantageous when small groups are involved. This is because the parameter estimates of small groups benefit from combining information from multiple groups within a cluster, leading to more accurate estimates, whereas, in case of ML-SEM, these estimates are affected by shrinking. We demonstrate the advantages of MixML-SEM through simulations and a real data example.

An early numeracy brief assessment: parametric and non-parametric IRT models

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. CECILIA MARCONI (Universidad de la República), Ms. Dinorah De León (Universidad de la República), Dr. Mario Luzardo (Universidad de la República), Dr. Alejandro Maiche (Universidad de la República)

This paper presents the development of a Uruguayan computerized maths test (Spanish acronym: PUMA). This test was devised to assess early numeracy competencies of preschool and first grade students. The importance of early mathematical skills as a basis for later development is well known. Many abilities such as approximate estimation, counting, or comparing numbers, among others, have been suggested as predictors of later mathematical achievement. The assessment of such competence is important in improving our understanding of the development of such abilities in order to assist children with difficulties early on. Nonetheless, maths assessment tools have been developed mostly in North American or European countries. Thus, they are not culturally adapted to or validated in Uruguay. The lengthy application process results in a gap in knowledge of mathematical skills for this population. PUMA consists of an online self-administered test screener involving nine subtests with 156 items in total. This study aims to develop a brief, psychometrically valid version of it. Item calibration was conducted with parametric models based on the IRT framework and kernel nonparametric regression IRT models. The participants were 443 preschool children aged 5 to 6 years old, a sample of whom were also given the standardized Early Mathematics Ability Test (TEMA-3). Considering this, the current article presents the development and comparison of the brief test version according to the parametric and non-parametric approach, and preliminary evidence of criterion validity. The results show that non-parametric models are as effective as parametric ones.

Latent Conjunctive Bayesian Network: Unify Attribute Hierarchy and Bayesian Network for Cognitive Diagnosis Modeling

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Seunghyun Lee (Columbia University), Dr. Yuqi Gu (Columbia University)

Cognitive diagnostic assessment aims to measure specific knowledge structures in students. To model data arising from such assessments, cognitive diagnostic models with discrete latent variables have gained popularity in educational and behavioral sciences. In a learning context, the latent variables often denote sequentially acquired skill attributes, which are often modeled by the so-called attribute hierarchy method. One drawback of the traditional attribute hierarchy method is that its parameter complexity varies substantially with the hierarchy's graph structure, lacking statistical parsimony. Additionally, arrows among the attributes do not carry an interpretation of statistical dependence. Motivated by these, we propose a new family of latent conjunctive Bayesian networks (LCBNs), which rigorously unify the attribute hierarchy method for sequential skill mastery and the Bayesian network model in statistical machine learning. In an LCBN, the latent graph not only retains the hard constraints on skill prerequisites as an attribute hierarchy, but also encodes nice conditional independence interpretation as a Bayesian network. LCBNs are identifiable, interpretable, and parsimonious statistical tools to diagnose students' cognitive abilities from assessment data. We propose an efficient two-step EM algorithm for structure learning and parameter estimation in LCBNs, and establish the consistency of this procedure. Application of our method to an international educational assessment dataset gives interpretable findings of cognitive diagnosis.

Do (non-)regularized partial correlation networks generalize? Re-evaluation of network generalizability

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Jeongwon Choi (Vanderbilt University), Dr. Alexander Christensen (Vanderbilt University)

The application of Gaussian graphical models (GGM) within the field of network psychometrics faces significant challenges concerning replicability. Some methods apply regularization to reduce overfitting due to the large number of parameters estimated. This study explores the use of (non-)regularization estimation methods for partial correlation networks to re-evaluate their generalizability.

Building on Williams & Rodriguez (2022), we investigate the role of (non-)regularization in enhancing the out-of-sample predictive ability of network models. We focus on two estimation methods for GGMs: the graphical least absolute shrinkage and selector operator with the extended Bayesian information criterion for model selection and a non-regularized maximum likelihood approach (Williams et al., 2019). Williams & Rodriguez (2022) found comparable to superior predictive performance for the non-regularized method but their approach was limited to three empirical datasets and deviated from best practices for ordinal data (Pearson rather than polychoric correlations), generalizability methods (leave-one-out cross-validation rather than *k*-folds cross-validation), and evaluation metrics (mean square error rather than accuracy).

Our study extends their investigation to a broader range of simulated conditions such as various data types and factor structures to better reflect real-world scenarios. We further re-evaluate Williams & Rodriguez's (2022) empirical examples using more appropriate correlations, generalizability methods, and evaluation metrics.

Contrary to existing arguments against regularized partial correlation networks, our findings demonstrate comparable or superior out-of-sample predictions and show evidence for reduced overfitting in small sample sizes. This study offers a more comprehensive view on how regularization improves network model generalizability.

Modeling age as predicting rater disagreement with a tri-factor model

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Linn Zapffe (Wake Forest University), Dr. Veronica Cole (Wake Forest University)

We often use multiple raters, such as multiple parents and teachers, in evaluations of children to get a more complete picture of the child's functioning and challenges. However, there are often differences among raters' assessments, both in terms of the overall level of the phenomenon under study and the nuances of how this phenomenon is measured. The latter can be considered a form of differential item functioning (DIF), whereby measurement parameters differ across children, but the presence of multiple raters per child creates the need for a measurement model that disentangles rater- and child-level variance. One such model is the *tri-factor model* (Bauer et al., 2013), which allows us to model differences between raters in the latent variable and the items used to measure it as a function of multiple rater- and child-level variables. In the current study, the tri-factor model is fitted to ratings on the Strengths and Difficulties Questionnaire from the Longitudinal Study of Australian Children (*N* = 9,538) which includes biannual data from children aged 4 to 16, with up to five raters per age group. The insights from the study could ultimately help uncover how rater disagreement changes with the age of the child and with that help inform decisions about the optimal number and type of raters to use in evaluations of children of different ages.

Asymptotic standard errors of equating coefficients using second-order delta method for non-parametric ability distribution

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Ikko Kawahashi (Meiji-Gakuin University), Dr. Yasuo Miyazaki (Virginia Polytechnic Institute and State University), Dr. Saori Kubo (Tohoku University)

One factor affecting the quality of scaled scores estimated using Item Response Theory (IRT) is the accuracy of equating. Therefore, the derivation of estimators for the standard errors of the equating coefficients has both theoretical and practical importance. Kawahashi (2023), under the setting of equating in a common-examinee design, derived estimators for the asymptotic standard errors of equating coefficients that do not require any distributional assumptions. This method has the advantage of being less biased than estimators that assume a normal distribution. As the assessment of equating accuracy is rigorous in the practical application of high-stakes testing, it is essential to assess the standard errors with even greater precision. This study used the second-order delta method to extend the asymptotic standard error estimator derived in Kawahashi (2023). In deriving this estimator, we referred to the asymptotic standard error estimator of indirect effects in path analysis models by Preacher et al. (2007). Moreover, an estimator of the asymptotic covariance between equating coefficients (A, B) was also derived to estimate the asymptotic standard error of the ability parameter. The simulation study showed that the proposed method did not improve the accuracy compared to the first-order delta method but provided a mathematical basis for constructing estimators using the higher-order delta method.

Cutoff for the Deleted-One-Covariance-Residual case influence measure in covariance structure analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Fathima Jaffari (Department of Tests and Measurement, National Center for Assessment, Education and Training Evaluation Commission (ETEC), Riyadh, Saudi Arabia), Dr. Jennifer Koran (Quantitative Methods Program, Southern Illinois University Carbondale, Carbondale, IL)

Influential cases in the data affect results from fitting models in covariance structure analysis. Case influence measures quantify the influence of each case on the modeling results. However, these case influence measures are typically model based measures, and their performance can be affected by error due to model misspecification. A new case influence measure, the Deleted-One-Covariance-Residual (*DOCR*; Jaffari & Koran, 2023), is a model-free alternative appropriate for use in covariance structure analysis. This study evaluated the adequacy of proposed cutoffs for the *DOCR*. The performance of *DOCR* in flagging simulated target cases was evaluated while the sample size, proportion of target cases to non-target cases, and type of model used to generate the data were manipulated. The mean of the 100 replications for the miss rates and their 95% confidence intervals were calculated using R package *psych* (Revelle, 2018). The results supported a recommended cutoff value of 0.008, the point at which no further reductions were observed in the miss rate.

Precision and Practicality: A Comparative Simulation of Measurement Properties of Visual Analog Scales and Likert Scales

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Siqi Sun (University of Virginia), Dr. Teague Henry (University of Virginia)

Accurately capturing the complexities of human experience is a central challenge for psychologists; yet, the choice of measurement scale can significantly impact data quality. This simulation study compared the performance of Visual Analog Scales (VAS) and Likert scales in capturing true latent scores under varying measurement error levels. Simulated data with varying parameters, including sample size, number of items, response categories (Likert), offset thresholds, and error levels, were generated. The performance of each scale was assessed using Spearman's correlation between true and measured scores. We found that VAS scales exhibited superior accuracy at low error levels (standard deviation ≤ 20) but declined significantly beyond that point. Well-validated Likert scales (with discrimination parameters ranging between 4.5-6.5) showed high accuracy across a wider error range, with minimal benefit from additional categories exceeding five. Three-point scales exhibited low accuracy and therefore were not recommended. Offsetting showed slight improvements in measurement precision for larger response options (VAS, 5, 7 pt Likert), and large improvements in the 3-point Likert scales. Increasing the number of items significantly improved performance for both VAS and Likert scales, while sample size had minimal impact for both measurements. Selecting the optimal scale hinges on anticipated error and the feasibility of employing validated Likert scales. Researchers are encouraged to prioritize validated items with low error, especially when utilizing single-item constructs common in Ecological Momentary Assessment (EMA) settings. Future research should further investigate the validity of VAS scales and report error measurement in addition to traditional indices of reliability or model fit.

Exploring Performance and Utilization of Generative AI Based Essay assessment

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Haeyeon Ahn (Ewha womans university)

This study explored the use of Generative AI technology to increase efficiency and provide meaningful feedback in essay assessment. The goal is to provide meaningful evaluation and feedback by lowering the teacher's time and cognitive burden in order to develop students' reading ability and thinking. The AI technology used here is an 'Chat GPT prompt engineering' that can be easily utilized in ordinary school situations. The basic principle of evaluating essays is to increase reliability and validity. The success of essay scoring depends on the scoring rubric and scoring guide. In addition, the evaluator should be valid and reliable. In this study, argumentation essays written by 10 students were evaluated by five evaluators and GenAI according to the same rubric. As a scoring method, an analytical scoring method was used, and the scoring rubric was pre-trained on the evaluator and GenAI. The reliability and validity of generative AI-based scoring were verified by comparing the results scored by the scorer with those scored by GenAI.

Psychometric properties of online social capital measurements in health studies: Scoping review

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Katarzyna Zawisza (Jagiellonian University Medical College), Michalina Gajdzica (Jagiellonian University Medical College), Paulina Sekuła (Jagiellonian University Medical College), Monika Brzyska (Jagiellonian University Medical College), Natalie Jagło (Jagiellonian University Medical College), Aleksandra Piłat-Kobla (Jagiellonian University Medical College), Paweł Jemioło (AGH University of Science and Technology), Dawid Storman (Jagiellonian University Medical College), Julia Ząber (Jagiellonian University Medical College), Małgorzata Bała (Jagiellonian University Medical College)

Although there is an intensive development of psychometry, sometimes the latest techniques are not used by researchers due to the complex mathematical apparatus used in a given technique. At the same time, the advantages of the new tools definitely support their use. In the last few decades, a number of techniques have been developed, as well as new approaches under the Item Response Theory (IRT), which are an alternative to the methods of the Classical Test Theory (CTT). The aim of the study was to summarize which methods and which aspects of validity and reliability were used during the development process of existing measurement tools of online social capital.

Theoretical or empirical studies concerning development or validation of the questionnaire methods to assess online social capital were considered. The studies conducted in the adult population from any country and sociocultural setting focused on health or directed to general population were checked. Formal query was carried out in the following databases: Medline (via Ovid), Embase, Web of Science, ProQuest, CINAH covering 1 January 2000 to 24 August 2023. 7040 studies were screened among which 81were classified as primary reference studies. Only in 44 papers the results of exploratory and/or confirmatory factor analysis were presented, few of them were built base on the IRT techniques. The psychometric properties of modified versions of the scales were rarely assessed. The study is funded in whole by National Centre of Science, Poland [2022/45/B/NZ7/04030].

Behavior of modified test statistics in correlation structure analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. You-Lin Chen (National Taiwan University), Dr. Li-Jen Weng (National Taiwan University)

Analysis of correlations among variables has long been of interest to the behavioral and social scientists (Bentler, 2007). Correlation structure analysis (RSA), similar to covariance structure analysis (CSA), enables the testing of complex correlational hypotheses by modeling the population correlation matrix directly (Bentler & Savalei, 2010). However, the test statistics obtained in RSA can be influenced by insufficient sample size and data nonnormality, akin to challenges in CSA (Fouladi, 2000). These distorted statistics can lead to problematic fit indices and a potentially misleading model evaluation. While several modified statistics from CSA can be extended to RSA, inferring their performance in RSA based on CSA findings is inappropriate due to the differences in statistical properties of sample correlations and covariances. To our knowledge, the performance of these modifications in RSA has not been extensively studied. We evaluated the empirical model rejection rates of seven modified statistics across five distributional conditions (normal, elliptical, skewed factor, skewed error, skewed both factor and error) and six sample sizes (150, 250, 500, 1,000, 2,500, 5,000). Two modifications for distribution-free statistics proposed by Yuan and Bentler (1997, 1999) failed under nonnormal conditions regardless of sample sizes. For normal-theory statistics, Browne's (1984) residual-based statistics performed poorly in all conditions except in large samples. Yuan-Bentler's (1998) two modifications for residual-based statistics and Satorra-Bentler's (1994) scaled statistics performed better but still failed when sample size was small and variables were nonnormal. Only Satorra-Bentler's adjusted statistics worked well in all conditions but tended to be slightly conservative.

Factor scores or sum scores when evaluating intervention effectiveness?

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Reeta Kankaanpää (University of Turku)

Randomized control trials (RCTs) are often used in psychology to make causal claims about the effects of interventions on adolescents' psychological wellbeing. An investigation of the psychometric properties of the assessment tools precedes the effectiveness evaluation. Ideally, intervention effects would be evaluated using the latent variable model that was deemed satisfactory in the psychometric evaluation. In practice, structural models involving latent variables and testing intervention effects are often too complex and have technical issues. Instead, intervention effects are estimated using sum scores where items are simply summed together. Sum scores may yield biased estimates and latent variable -based factor scores could produce much better estimates for intervention effects. However, less is known about the superiority of factor scores over sum scores in the case where the measures might contain systematic error in addition to random error. This study will investigate the effect of scoring method on intervention effects using simulation and empirical analysis. We will evaluate two scenarios: with and without a possible systematic error, termed as method-related effect. First, with simulated data, we will compare the estimates from factor scores and sum scores to estimates from a latent variable model when testing the intervention effect. Second, with empirical data, we will compare the estimates from factor scores and sum scores with a real school intervention study. This study will guide researchers conducting RCTs whether to replace sum scores with factor scores in future studies.

Proposing a two-step simulation procedure to assess replicability for planned studies

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Denny Kerkhoff (Bielefeld University)

With replication rates of psychological research findings remaining unsatisfactory, replicability research as a meta-research field is at the forefront of securing credibility of scientific research. A prerequisite for successful replication is ensuring that the planned primary study is suited to detect a target population effect with sufficient statistical power and estimates that are precise and unbiased. A prominently suitable method to assess these statistical properties for a planned study is the (Monte Carlo) simulation study. Simulations with the goal to infer under what conditions the planned primary study uncovers assumed effects in the population have become an established tool in psychological research, especially in the form of a-priori power analyses to determine required sample sizes. However, simulations can additionally be set up to simulate under what conditions potential future replication studies recover plausible estimates of a planned primary study. Conceptualizing simulation studies in such a way enables researchers to a-priori assess credible replication rates for a study. However, at present, there is no established simulation workflow to obtain such estimated replication rates. As a result, simulation capabilities to address replicability concerns are currently underutilized. In this presentation, I first provide a short overview on how and to what extend replicability of planned research projects is currently evaluated in psychological research, based on reports in published research and preregistrations. Then, I present a proposed two-step simulation framework designed to assess and optimize estimated replication rates of a planned statistical analysis.

Comparing regularization, alignment, and model modification for partial measurement invariance

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Emma Somer (McGill University), Dr. Carl Falk (McGill University), Dr. Milica Miočević (McGill University)

When seeking a partial measurement invariance model, many traditional approaches involve an iterative procedure, where items are tested for noninvariance one at a time, and subsequently, latent parameter estimates are interpreted. More recently, several approaches that simultaneously identify noninvariant items and perform latent parameter estimation have been proposed. In particular, regularization methods involve imposing a penalty term that shrinks parameter differences across invariant items to zero in order to reduce model complexity and improve interpretation. In addition, following a configural model, alignment finds latent means and co-variances that minimize a criterion representing measurement noninvariance. In the current study, we compare lasso, ridge, and elastic net with alignment and traditional multiple group CFA for estimating latent parameters in a simulation study. The aim of this research is to 1. Compare traditional approaches to measurement invariance testing to regularization methods, and 2. Evaluate whether the methods can produce unbiased and efficient estimates of the correlation between two latent variables. We manipulated the number of indicators (4 and 8), sample size (N = 200, 500, 1000), magnitude of noninvariance (small, medium, and large), proportion of noninvariance (25%, 50%, and 75%), and the value of the correlation (0 and 0.3) and tested a range of penalty values. Preliminary results indicate that ridge regression produces less efficient and biased point estimates when the proportion of noninvariance is large, whereas alignment generally outperforms the methods under the same conditions.

Using Machine Learning for Extracting Personality Insights from Projective Responses on Inkblots

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Krystof Petr (Department of Psychology, Faculty of Arts, Palacký University Olomouc), Dr. Daniel Dostál (Department of Psychology, Faculty of Arts, Palacký University Olomouc)

The projective hypothesis posits that the entirety of human behavior can serve as a window into personality. Traditional projective techniques, particularly those prompting verbal responses to stimuli such as inkblots, have been pivotal in exploring this hypothesis. Despite their popularity and potential, these methods face criticisms regarding validity, reliability, and cost-effectiveness. Addressing these concerns, our study explores the feasibility of employing machine learning algorithms to accurately deduce personality traits from responses on custom-made projective materials.

In our investigation, over 5,000 participants provided verbal responses to inkblots. Initially, we analyzed these responses for the presence of ten content categories, which were hypothesized to correlate with the Big Five personality traits and Positive and Negative Implicit Affect. Subsequently, employing a variety of regression and classification techniques, including neural networks, we examined the ability of these machine learning models to predict personality characteristics based on quantitative analyses of the responses, transformed into vector space via word embeddings.

Our findings illuminate the personality-relevant information in inkblot responses, revealing both anticipated and novel associations. The most effective machine learning models demonstrated a capability to extract a considerable volume of personality insights from the embedded representations of responses, achieving accuracy comparable to analyses conducted with social network data. These results underscore the depth of personality insights in projective stimuli and suggest promising future research and development directions.

Topic Analysis of the AI-based Assessment using LDA Topic Modeling

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Hun Won Choi (Ewha womans university), Prof. Youn-Jeng Choi (Ewha womans university)

Efforts to integrate and utilize AI in education are expanding globally. In some countries, these efforts have accelerated significantly post-COVID-19, especially in the area of educational assessment. This study aims to explore the trends in AI-based assessment globally by analyzing prevalent topics of discussion within the research field and examining current research directions and challenges. A total of 979 education-related papers published until February 2024 were analyzed using the R program (version 4.3.1) to conduct LDA topic modeling. From 968 analyzable papers, 6,297 words were extracted, and the number of topics was determined to be seven through a hyperparameter tuning process, considering the ease of interpretation and the criteria by Deveaud et al. (2014) and Griffiths and Steyvers (2004). The analysis included frequency analysis of appearing words, word clouding, LDA topic modeling, and calculation of word frequency and significance using TF and TF-IDF. The extracted topics included Information and Communication Technology (ICT), online education platforms like MOOCs, medical education, Natural Language Processing (NLP) applications like chatbots and ChatGPT, and considerations of validity and fairness in assessment performance. These topics revealed that research related on test validity and fairness was actively conducted to introduce and utilize AI in the field of educational assessment. Additionally, a significant portion of research focused on NLP technologies, including chatbots and ChatGPT. Future presentations will delve into how research objectives evolve over time in the educational field regarding AI-based assessment, aiming to discuss critical issues surrounding AI-based assessment.

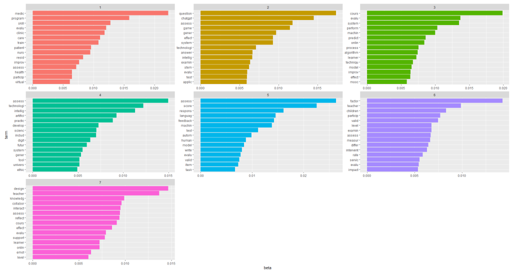


Figure 2. keyword frequency rate by topics for ai-based assessment.png

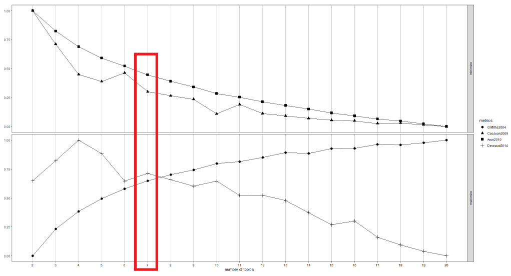


Figure 1. model performance graph by hyperparameter tuning for determining the number of topics for ai-based assessment.png

Military Stigma Scale invariance across National Guard education and paygrade

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Sam Cacace (Violence Prevention Center, University of North Carolina at Charlotte), Dr. Robert Cramer (Violence Prevention Center, University of North Carolina at Charlotte), Mr. Max Stivers (Department of Psychology, Louisiana State University), Dr. Raymond Tucker (Department of Psychology, Louisiana State University), Dr. Marcus VanSickle (Themis Forensic Psychological Services)

The Military Stigma Scale (MSS) is a 26-item measure containing two subscales: Public Stigma and Private Stigma (Skopp, et al., 2012), and found to conform to a bi-factor measurement structure with a General Stigma factor, and a Methods specific factor for reverse-coded items (Vidales, et al., 2021). In the current study, scalar invariance was evaluated in a sample of n = 1832 Army National Guard members across paygrade levels (junior n = 848; senior n = 402; officer n = 279) and education levels (no college degree n = 1.116; college degree n = 711). Low expected common variance (ECV) was found for the Public Stigma specific factor for junior and senior enlisted (ECV = 5.1%) compared to officers (ECV = 24.8%), and for those without a college degree (ECV = 5.0%) compared to those with a college degree (ECV = 40.3%). After scalar invariance, significant relative factor mean differences were found for Private Stigma among senior enlisted (ΔM = -0.36, p < .001) and officers (ΔM = -0.39, p< .001) as compared to junior enlisted, and Public Stigma was significantly higher in officers (ΔM = 0.27, p = .030). For college degree holders, Public Stigma (ΔM = 0.22, *S.E.* = 0.08, p = .006) and General Stigma (ΔM = 0.12, p < .001) were both significantly higher than in those without a college degree, and Private Stigma was significantly lower in college degree holders (ΔM = -0.27, p < .001) than in those without a college degree.

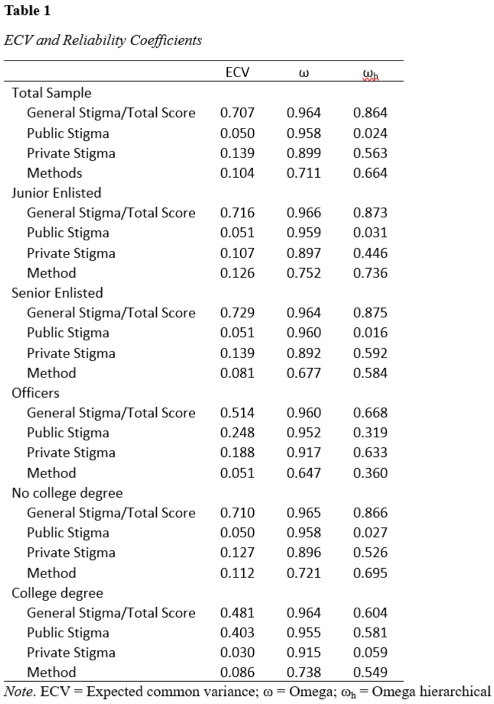


Table 1

ECV and Reliability Coefficients

	ECV	ω	ω_h
Total Sample			
General Stigma/Total Score	0.707	0.964	0.864
Public Stigma	0.050	0.958	0.024
Private Stigma	0.139	0.899	0.563
Methods	0.104	0.711	0.664
Junior Enlisted			
General Stigma/Total Score	0.716	0.966	0.873
Public Stigma	0.051	0.959	0.031
Private Stigma	0.107	0.897	0.446
Method	0.126	0.752	0.736
Senior Enlisted			
General Stigma/Total Score	0.729	0.964	0.875
Public Stigma	0.051	0.960	0.016
Private Stigma	0.139	0.892	0.592
Method	0.081	0.677	0.584
Officers			
General Stigma/Total Score	0.514	0.960	0.668
Public Stigma	0.248	0.952	0.319
Private Stigma	0.188	0.917	0.633
Method	0.051	0.647	0.360
No college degree			
General Stigma/Total Score	0.710	0.965	0.866
Public Stigma	0.050	0.958	0.027
Private Stigma	0.127	0.896	0.526
Method	0.112	0.721	0.695
College degree			
General Stigma/Total Score	0.481	0.964	0.604
Public Stigma	0.403	0.955	0.581
Private Stigma	0.030	0.915	0.059
Method	0.086	0.738	0.549

Note. ECV = Expected common variance; ω = Omega; ω_h = Omega hierarchical

Table1 ecv.png

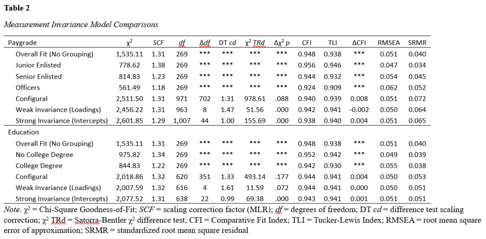


Table 2

Measurement Invariance Model Comparisons

Paygrade	χ^2	*SCF*	*df*	Δ*df*	DT *cd*	χ^2 *TRd*	$\Delta\chi^2$ *p*	CFI	TLI	ΔCFI	RMSEA	SRMR
Overall Fit (No Grouping)	1,535.11	1.31	269	***	***	***	***	0.948	0.938	***	0.051	0.040
Junior Enlisted	778.62	1.38	269	***	***	***	***	0.956	0.946	***	0.047	0.034
Senior Enlisted	814.83	1.23	269	***	***	***	***	0.944	0.932	***	0.054	0.045
Officers	561.49	1.18	269	***	***	***	***	0.924	0.909	***	0.062	0.052
Configural	2,511.50	1.31	971	702	1.31	978.61	.088	0.940	0.939	0.008	0.051	0.072
Weak Invariance (Loadings)	2,456.22	1.31	963	8	1.47	51.56	.000	0.942	0.941	-0.002	0.050	0.064
Strong Invariance (Intercepts)	2,601.85	1.29	1,007	44	1.00	155.69	.000	0.938	0.940	0.004	0.051	0.065
Education												
Overall Fit (No Grouping)	1,535.11	1.31	269	***	***	***	***	0.948	0.938	***	0.051	0.040
No College Degree	975.82	1.34	269	***	***	***	***	0.952	0.942	***	0.049	0.039
College Degree	844.83	1.22	269	***	***	***	***	0.942	0.930	***	0.055	0.038
Configural	2,018.86	1.32	620	351	1.33	493.14	.177	0.944	0.941	0.004	0.050	0.053
Weak Invariance (Loadings)	2,007.59	1.32	616	4	1.61	11.59	.072	0.944	0.941	0.000	0.050	0.051
Strong Invariance (Intercepts)	2,077.52	1.31	638	22	0.99	69.38	.000	0.943	0.941	0.001	0.051	0.051

Note. χ^2 = Chi-Square Goodness-of-Fit; *SCF* = scaling correction factor (MLR); *df* = degrees of freedom; DT *cd* = difference test scaling correction; χ^2 TRd = Satorra-Bentler χ^2 difference test; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual

Table2 modelcomparisons.png

A holistic view of academic performance: Beyond Averages with MELSM and Spike-and-Slab

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Marwin Carmo (University of California, Davis), Dr. Donald Williams (University of California, Davis), Prof. Philippe Rast (University of California, Davis)

Typically, research on academic performance centers on the average academic achievement of a student or a school, but this measurement does not fully capture the learning experience. To get a more complete picture, it's important to look at both average performance and variability within a cluster. Our approach involves identifying clusters with unusually large or small within-cluster variance in academic achievement, which can indicate inconsistent or consistent performance. We adapted the mixed-effects location scale model (MELSM) using the Spike and Slab regularization technique to shrink random effects to their fixed effect. This approach allows us to identify clusters with unusually (in)consistent academic achievement. We focus on identifying clustering units with unusually high or low consistency (residual variance) in academic achievement. The Spike and Slab prior serve as the Bayesian analog to lasso regularization. However, it allows one to combine it with Bayes factors that provide a decision boundary for the identification of (in)consistent clustering units. Our approach provides a more nuanced understanding of academic performance by considering average scores and variability within clusters, which can help educators provide more targeted support to students and schools.

The crossed random effects drift diffusion model – a simulation study

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Nele Bögemann (University of Münster), Prof. Steffen Nestler (University of Münster)

To properly capture interindividual variability in cognitive processes, cognitive modelers increasingly employ hierarchical Bayesian models in which subjects are treated as random effects. Additional random effects may be added to the model for stimuli, for example, when subjects are crossed with social target stimuli as in social cognition experiments (Judd et al., 2012). To date, few simulation studies have comprehensively investigated the estimation performance of such more complex hierarchical cognitive models. In our simulation study, we sought to close this gap for the crossed random effects variant of the Drift Diffusion Model (DDM; Ratcliff, 1978; Vandekerckhove et al., 2010). We used a simulation design with two crossed random effects - mirroring subjects and targets as in social cognition experiments - and we varied design settings in ways realistic to such experiments. Specifically, we manipulated the variance of subject and target population distributions, mirroring homo- vs. heterogeneous populations, as well as the number of draws from each population, mirroring subject and trial numbers. Additionally, we manipulated model complexity by inducing constraints on the estimated random effect structure (crossed vs. single vs. no random effects). All models were estimated in JAGS and their estimation was evaluated based on different performance criteria (e.g., bias). Importantly, performance evaluation considered both the individual and the population level for both subjects and targets, providing novel insights into the interplay of multiple random effects on cognitive parameter estimation across levels.

Item and Test Development for Next Generation Assessments

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Eunji Lee (University of Georgia), Prof. George Engelhard (University of Georgia)

This study examines two approaches for addressing some of the innovations posed by next generation assessments. The two specific approaches to item and test development that guide this study are Assessment Engineering (AE) and the Constructing Measures Framework (CMF). Some of the important innovations of next generation assessments include Automatic Item Generation, Automated Test Assembly, Computer-Based Testing, and Computer-Adaptive Testing. Each of these innovations pose challenges for item and test development. AE is a comprehensive framework for designing and efficiently building scales using technology-driven item writing and test development. AE emphasizes several critical processes including construct mapping and evidence modeling, task model building, and psychometric calibrations. CMF includes four key building blocks: the definition of a latent variable, development of an observational design, use of a scoring system to represent the observational design, and a measurement model. Rasch measurement theory is used as the fourth building block in this study to connect the latent variable, observations, and scoring rules.

AE and CMF have some overlapping, as well as unique aspects, and these are discussed in this study. The purpose of this study is to explore the integration of next generation measurement concepts using AE and CMF. The overall goal is to support psychometrically sound item and test development. These innovations hold the promise of improving the appropriate use and interpretation of scores on educational and psychological assessments.

Key words: Item and test development, next generation assessments, assessment engineering

Innovation as a Dynamic System: A Network Approach to Employee Innovativeness

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mrs. Kamila Zahradnickova (Prague University of Economics and Business)

Innovation plays a crucial role in driving social progress, and understanding employee innovativeness is essential for organizations seeking to enhance their performance and gain a competitive advantage. This research paper explores the concept of employee innovativeness using network analysis (Borsboom & Cramer, 2013) and builds upon the integrative measure of innovativeness developed by Lukes and Stephan (2017). This measure treats innovativeness as a system and accounts for both individual and contextual factors measured in two scales, Innovative Behaviour Inventory (IBI) and Innovation Support Inventory (ISI). While Lukes and Stephan employed the common factor model (CFM) framework to create a measure of employee innovativeness, this paper proposes an alternative approach using network analysis. Networks could provide a more nuanced understanding of the internal structure of innovativeness, revealing interconnectedness, emergent properties, and reciprocal relationships among individual items. Re-analysing a dataset of 2812 participants originally collected by Lukes and Stephan, three networks were estimated to confirm dimensionality and uncover new questions for future research. Two separate network models supported the dimensionality of Lukes and Stephan's measures IBI and ISI. Interestingly, the third network with items from both IBI and ISI revealed little connection between contextual and individual factors, which raises questions for future research exploring how companies may promote innovative behaviour. Overall, networks offer an alternative perspective on employee innovativeness and may enable the development of new interventions to increase employee innovativeness.

A signal cancelling approach to exploratory factor analysis.

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. André Achim (Université du Québec à Montréal)

Signal cancelling provides a radically new and efficient approach to exploratory factor analysis, without matrix decomposition. Suitable contrasts of pairs of unifactorial indicators of the same factor can cancel the factor information. This leaves only noise in such contrasts, implying null correlations with all remaining variables. Contrasts of indicators of different factors cannot cancel the signal and therefore retain non-null correlations with some other variables. The contrast weights are obtained by minimizing the sum of squares of these v-2 correlations. Multiplying this criterion by N-1 provides a X^2 with v-2 degrees of freedom to assess signal cancellation. With reasonable sample sizes, X^2 significance tells unsuccessful from successful signal cancellation, the latter characterizing indicator pairs that correspond to one of the factors. Combinations of more variables can similarly cancel the signal of multifactorial indicators. Factor loadings are obtained from the optimized contrast weights, their correlations from those of pooled unifactorial indicators. Signal cancellation has other attractive properties, such as no need to decide on number of factors, no risk of Heywood cases, no indeterminacy of doublet factor loadings; even unique variables are readily set aside.

A Study of Maternal Involvement, Family Environmental Diversity and Social Competence of Young Children in Different Ethnic Groups-Secondary Data Research from Kids in Taiwan-National Longitudinal Study of Child Development and Care (KIT)

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Prof. Chia-Yen Hsieh (National Pingtung University), Prof. Chiu-Hsia Huang (National Pingtung University)

Using a secondary data analysis by parent questionnaire data for 36-month-old children from Kids in Taiwan (KIT) project – a national longitudinal study of child development and care, this study aimed to analyze the differential effects of maternal involvement, family environmental diversity, and social competence for 2,164 young children in different ethnic groups of mothers. With a cluster analysis based on the variables of maternal involvement and family environmental diversity, two groups were divided: high involvement and low involvement. Subsequently, a two-factor multivariate analysis was conducted separately for each ethnic group. The significant effects of maternal involvement and family environmental diversity for Taiwanese and Chinese mothers, while only family environmental diversity had a significant effect for Southeast Asian mothers. No significant difference was for Taiwanese Indigenous mothers. Post hoc comparisons revealed that higher levels of involvement among Taiwanese mothers were associated with higher levels of independence, proactivity, sociability, and obedience in children, however, higher levels of involvement among Chinese mothers were only associated with increased children's obedience. Regarding family environmental diversity, higher levels in Taiwanese mothers were associated with higher levels of independence, proactivity, sociability, and obedience in children. Higher levels in Chinese mothers were linked to increase independence and proactivity in children, while higher levels in Southeast Asian mothers were associated with increased proactivity and sociability in children. Consequently, both maternal involvement and family environmental diversity independently influenced children's social competence development, but their effects vary among different ethnic groups of mothers. Findings suggest new directions for future research.

DIF detection using Rasch trees method and mixture Rasch model

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Sugyung Goh (Ewha womans university), Prof. Youn-Jeng Choi (Ewha womans university)

This study aims to examine how the Rasch trees method and the mixture Rasch model detect Differential Item Functioning (DIF) differently using their own groups in the PISA 2022 mathematics assessment. The Rasch trees method allows DIF detection based on model-based recursive partitioning and conducts statistical validations to identify DIF, whereas the mixture Rasch model detects DIF among previously unknown (latent) groups (Strobl, Kopf, & Zeileis, 2015; Rost, 1990). These two approaches do not have pre-specified groups, but the methods they use to identify groups differ from each other.

A preliminary investigation was conducted using data from South Korean students who responded to the M04 mathematics item set. A total of 29 survey items related to students' backgrounds, mathematical characteristics, and other relevant factors were selected to detect groups of subjects exhibiting DIF based on literature reviews. The Psychotree R package was utilized for Rasch trees analysis (Zeileis et al., 2022), revealing the existence of DIF due to two explanatory variables: studying for school or homework before or after school and parents' expectations for their child's career (see Figure 1).

We will perform a mixture Rasch model analysis to detect latent group DIF using WINMIRA 2001 (von Davier, 2001), and multinomial logistic regression will be used to identify the characteristics of latent groups using student survey items. The utilization of two methods based on the Rasch model for comparing and analyzing DIF is expected to offer diverse perspectives, leading to more valid and reliable results.

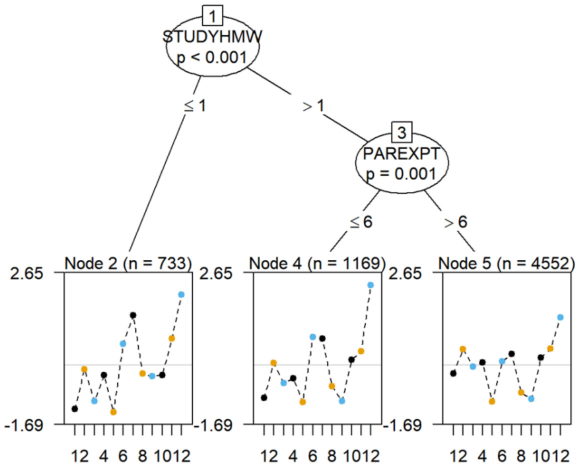


Figure 1. DIF Detection by STUDYHMW and PAREXPT Variables using Rasch Trees

Figure 1. dif detection by studyhmw and parexpt variables using rasch trees.png

Research trends in alternative education using keyword network analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Seongkyung Kim (Ewha womans university), Prof. Youn-Jeng Choi (Ewha womans university)

This study aims to comprehensively understand trends in alternative education and alternative school research through keyword network analysis. Alternative education/schools, distinguished by their innovative learning methods and educational systems, encompass student-centric approaches, personalized learning paths, an emphasis on creativity, and the enhancement of problem-solving skills. The necessity of alternative education has been acknowledged globally, with various forms existing for a considerable time and consistently attracting attention, particularly in countries like Canada, the UK, the USA, Germany, Japan, India, Taiwan, etc. In South Korea, the escalating competition in entrance exams has sparked a recent surge in interest and participation in alternative education. Consequently, it becomes crucial to compare and analyze the research trends in alternative education/schools both domestically and internationally. To achieve this, a keyword network analysis will be executed, involving techniques such as word extraction, calculation of co-occurrence frequency, evaluation of word similarity, and network formation (Lee, 2021), building upon the research findings of Girvan and Newman (2002), Clauset et al. (2004), and Newman (2006). We will extract and refine author keywords related to alternative education/schools from academic databases such as SCI, SSCI, SCOPUS, and KCI. In the detailed analysis, we will perform centrality analysis, and clustering analysis using Netminer 4.5. The centrality analysis will involve examining degree centrality and betweenness centrality, while clustering analysis will utilize convergence of iterated correlations. This research aims to make a significant contribution to the global advancement of alternative education/schools by providing insights into research trends both domestically and internationally.

The Response Style in Continuous Bounded Response: Modeling and Application

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Youxiang Jiang (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China), Prof. Hongbo Wen (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China)

While many models like item response tree (IRTree) have been developed to analyze response styles in Likert scales, the same attention has not been given to questionnaires with continuous measurement formats. As computer-based measurements become increasingly popular, continuous bounded response data are becoming more prevalent in questionnaire data. These data sources encompass elements such as the visual analogue scale (VAS), slide bars, probability judgments, etc. In response, we developed an item response model framework for response styles in continuous bounded responses. Our model framework, which is based on a hierarchical structure and builds pseudo responses, can flexibly incorporate content traits, extreme responses, and midpoint response styles. To validate this new model for evaluating response styles, we conducted an empirical study, in which the Likert and VAS versions of three questionnaires have been analyzed. The results showed that the new model fits the continuous response best (as shown in Table 1), and the content traits and response style tendencies estimated by the new model highly correlate with those estimated by the IRTree model (as shown in Figure 1). Furthermore, a simulation study has been conducted to test the recovery of model parameters under various situations. The results demonstrated that all parameters can be accurately estimated using the Markov chain Monte Carlo method. Overall, the content traits and response styles estimated by the new model exhibit high validity, our model effectively addressing the gap in response styles for continuous bounded response data.

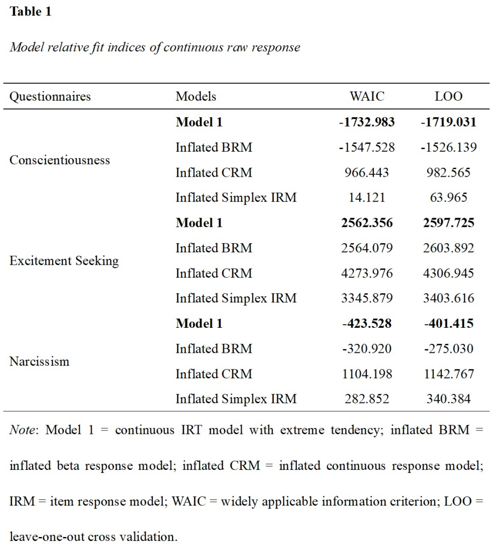


Table 1

Model relative fit indices of continuous raw response

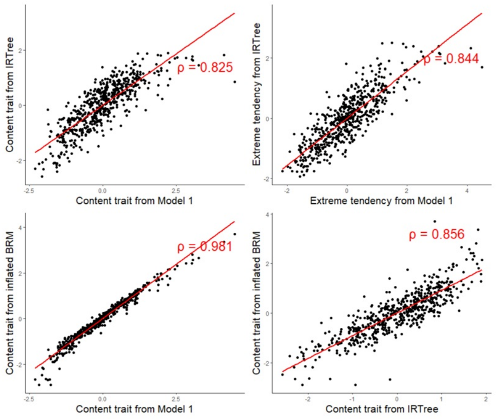
Questionnaires	Models	WAIC	LOO
Conscientiousness	**Model 1**	**-1732.983**	**-1719.031**
	Inflated BRM	-1547.528	-1526.139
	Inflated CRM	966.443	982.565
	Inflated Simplex IRM	14.121	63.965
Excitement Seeking	**Model 1**	**2562.356**	**2597.725**
	Inflated BRM	2564.079	2603.892
	Inflated CRM	4273.976	4306.945
	Inflated Simplex IRM	3345.879	3403.616
Narcissism	**Model 1**	**-423.528**	**-401.415**
	Inflated BRM	-320.920	-275.030
	Inflated CRM	1104.198	1142.767
	Inflated Simplex IRM	282.852	340.384

Note: Model 1 = continuous IRT model with extreme tendency; inflated BRM = inflated beta response model; inflated CRM = inflated continuous response model; IRM = item response model; WAIC = widely applicable information criterion; LOO = leave-one-out cross validation.

Table 1.jpg

Figure 1

The correlation between latent traits estimated by Model 1 and other models for conscientiousness



Note: Model 1 = continuous IRT model with extreme tendency; IRTree = item response tree model; inflated BRM = inflated beta response model; Model 1, inflated BRM, and IRTree used to fit continuous raw response, continuous raw response and pseudo response, and Likert response, respectively.

Figure 1.jpg

How to include dichtomous variables in Meta-Analytic Structural Equation Modeling?

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Hannelies de Jonge (University of Amsterdam), Belén Fernández-Castilla (National Distance Education University), Kees-Jan Kan (University of Amsterdam), Frans J. Oort (University of Amsterdam), Suzanne Jak (University of Amsterdam)

Meta-analytic structural equation modeling (MASEM) is a method to systematically synthesize results from primary studies, allowing the researchers to simultaneously examine multiple relations among variables by fitting a structural equation model to the pooled correlations. An advantage of MASEM is that one can include effect sizes of primary studies even when they do not include all variables of interest. However, incorporating dichotomous variables (e.g., having a disease or not) into MASEM poses challenges. While primary studies that investigate the relation between a dichotomous and continuous variable typically report standardized mean differences (e.g., Cohen's *d*), MASEM requires correlation matrices as input. This can be solved by converting standardized mean differences to point-biserial correlations. However, here lies a complication because, in contrast to a standardized mean difference, the point-biserial correlation depends on the distribution of group membership (e.g., the prevalence of a disease). In this work, we will discuss and evaluate various conversion methods (i.e., different formulas and different ways to accommodate distribution dependencies). Using simulated data, we investigate the effects of these conversion methods on the estimation of the MASEM parameters. In our simulation study, we also vary the sampling method used in primary studies, the population "base rate" (e.g., the prevalence), the within-study sample size, and the distribution of respondents over two groups. We present our preliminary results and introduce a user-friendly web application for converting primary study statistics into MASEM-compatible effect size.

A note on the bias correction of estimated eigenvalues applied to a correlation matrix, with applications to factor analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Christopher Haverly (University of Hawaii at Manoa), Kentaro Hayashi (University of Hawaii at Manoa)

In high-dimensional data, estimated eigenvalues are well-known to exhibit bias compared to their population counterparts, particularly in the leading eigenvalues. To address this bias, Shen, Shen, Zhu, and Marron (2016) introduced a correction formula, later refined by Wang and Fan (2017). The latter authors applied this bias correction to the factor analysis model, naming their method the S-POET (Shrinkage Principal Orthogonal complEment Thresholding) method. Through simulations, they demonstrated that S-POET reproduced the population covariance matrix more accurately than the sample covariance matrix, assessed using three norms (spectral, Frobenius, and max norms). However, in the fields of psychology and other social sciences, factor analysis is commonly performed on a correlation matrix rather than a covariance matrix. While the correlation matrix is a derivative of the covariance matrix, it remains unclear whether results obtained from the covariance matrix can be directly translated to the correlation matrix. The performance of S-POET using a correlation matrix has not been thoroughly investigated. The current research aims to bridge this gap.

The Impact of Measurement Non-Invariance on ANCOVA Performance

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. You Kyoung Hwang (Sogang University), Prof. Hye Won Suk (Sogang University)

The Analysis of Covariance (ANCOVA) serves as a cornerstone in evaluating treatment effects within randomized pretest-posttest control group design studies. Traditionally, ANCOVA operates under the assumption of measurement invariance across groups and time points, typically relying on aggregate scores of observed variables. However, intervention studies often encounter response shift phenomena, altering item interpretation and inducing measurement non-invariance at the posttest across groups. Such non-invariance can introduce bias into ANCOVA results. This study delves into the ramifications of measurement non-invariance on ANCOVA's efficacy in estimating treatment effects. Through a systematic simulation study, we investigate the effects of varying levels of intercept non-invariance, the proportion of non-invariant items, and sample size on ANCOVA's performance. We assess outcome measures including bias, MSE, coverage, and power (or type I error) to evaluate ANCOVA's performance. Additionally, we compare ANCOVA's performance under various types of measurement invariance, including non-invariance on loadings and non-invariance across groups at the pretest, as well as different models utilizing latent variables. Our findings illuminate the susceptibility of ANCOVA to measurement non-invariance and provide insights into its implications for intervention study design and analysis.

Predictors of response-time effort across countries in PISA 2015 science

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Michalis Michaelides (University of Cyprus), Dr. Demetris Avraam (University of Liverpool), Ms. Militsa Ivanova (University of Cyprus)

Examinees' test-taking effort in achievement tests has a significant impact on their test outcomes (Wise & DeMars, 2005). In low-stakes assessment programs, with few or no individual consequences, test-takers may not be motivated to invest adequate effort, and thus their test scores may underestimate their true proficiency. The current study quantifies test-taking effort using item response times from the PISA 2015 Science assessment and examines individual predictors of the response time effort (RTE) index. Data were obtained from 56 countries participating in the computerized administration of the program. Country sample sizes of 15-year-old examinees ranged between 3371 and 23141. Rapid guessing (and rapid omission) behavior was quantified with the 15% normative threshold: responses provided before the 15% of the mean item response time were flagged as rapid guessing. Examinee indices for RTE were calculated as the proportion of item-level responses for which an examinee did not exhibit rapid guessing behavior (Wise & Kong, 2005). Country average RTE scores ranged from 89.9% to 98.7%. A multilevel regression model accounting for school clustering will be fitted on each country sample to estimate the effects of demographic, achievement, and motivational variables as individual-level predictors of RTE. Country-specific estimates will then be combined to overall estimates across the 56 country samples through random-effects meta-analysis. The findings will be useful in identifying robust predictors of effortful test-taking that generalize cross-culturally or appear to be country-specific. Targeted interventions can then be designed to discourage rapid guessing in low-stakes assessments.

Comparison of component-based SEM methods in testing component interaction effects

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Zhiyuan Shen (McGill University), Dr. Gyeongcheol Cho (The Ohio State University), Dr. Heungsun Hwang (McGill University)

Structural equation modeling (SEM) has two distinct domains–factor-based and component-based, depending on whether a construct is statistically represented as a (common) factor or a component or weighted composite of observed variables. Generalized structured component analysis (GSCA) and partial least squares path modeling (PLSPM) are full-fledged methods for component-based SEM. They involve different approaches to testing the interaction effects of components. GSCA provides a single-step approach that allows the specification of a single model with all the main and interaction effects of components and the simultaneous testing of the effects. In contrast, PLSPM's primary approach carries out two steps sequentially. In the first step, it fits a model with only the main effects of components, estimating the components. In the second, it fits a path analytic model using the components as indicators and estimates their main and interaction effects. Despite such technical differences between the approaches, no study has examined their performance in testing the interaction effects of components. To fill this gap, we conducted a simulation study to systematically evaluate the performance of GSCA and PLSPM in testing component interaction effects under various experimental factors, such as measurement model specification, sample size, and effect size.

Response styles stability modelled with IRTrees

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Marek Muszyński (Institute of Philosophy and Sociology of the Polish Academy of Sciences), Dr. Tomasz Żółtak (Institute of Philosophy and Sociology of the Polish Academy of Sciences), Prof. Artur Pokropek (Institute of Philosophy and Sociology of the Polish Academy of Sciences)

Response styles (RS) time stability was evidenced (e.g., Weijters et al., 2010; Wetzel et al., 2016) but not with the use of newly proposed models, i.e., IRTrees (Boeckenholt, 2012; Khorramdel & von Davier, 2014) and multidimensional generalized partial credit models (Henninger & Meiser, 2019). Moreover, it remains unclear whether response styles have the same time stability in different survey scales.

We aim to investigate the RS time stability using newly developed longitudinal IRTree models (Ames, 2022; Ames & Leventhal, 2021), using data from two web survey studies conducted on two occasions, separated by ca. 14 days on a group of participants recruited from an opt-in panel. The surveys lasted around 20 minutes. Study 1 (N= 401 participants) comprised measures of reading behaviour and vaccine attitudes, measured on a 4-point rating scale.

We employed multidimensional IRTree models to estimate RS time stability and to study cross-scale RS time stability and its covariates (gender, age, survey experience, and self-reported education level). The results favoured the scale-specific RS model over the general RS model, pointing to different response processes in different measures, although cross-scale RS correlations amounted to 0.44-0.83. Inter-measurement correlation of RS (0.77) pointed to a considerable RS time stability. Gender and education correlated with the RS level but not its change over time. The Study 1 conclusions were expanded in Study 2 (N = 485), where similar analyses were run, but with the use of different scales (personality, trust, empathy, altruism, social norms) and a different rating scale (5-point).

Sensitivity of goodness-of-fit indices to model misspecification in Bifactor models: A simulation study examining the consequences of ignoring cross-loadings.

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Carmen Ximenez (Universidad Autónoma de Madrid), Dr. Javier Revuelta (Universidad Autónoma de Madrid), Mr. Cesar Piris (Universidad Autónoma de Madrid)

Bifactor latent models have gained popularity and are widely used to model construct multidimensionality. When adopting a confirmatory approach, a common practice is to assume that all cross-loadings take zero values. This poster presents the results of a simulation study that explores the impact of ignoring non-zero cross-loadings on the performance of confirmatory bifactor analysis. The study assesses the sensitivity of goodness-of-fit indices in detecting model misspecification resulting from ignoring non-zero cross-loadings. Several commonly used structural equation modeling (SEM) fit indices are examined, including both biased estimators of the fit index (CFI, GFI, and SRMR) and unbiased estimators (RMSEA and SRMR$_u$). Our results indicate that commonly used SEM fit indices are not useful in detecting model misspecifications due to ignoring moderate and large cross-loading values and using small simple sizes. We recommend the use of the unbiased SRMR index (SRMR$_u$) with a cutoff value adjusted by the communality level (R^2), as it is the only fit index sensitive to model misspecification resulting from ignoring non-zero cross-loadings in the bifactor model. The findings of this study provide insights into modeling cross-loadings in confirmatory bifactor models and offer practical recommendations to researchers.

Overfactoring in skewed rating scale data: A comparison between factor analysis and the graded response model.

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Javier Revuelta (Universidad Autónoma de Madrid), Dr. Carmen Ximenez (Universidad Autónoma de Madrid), Ms. Noelia Yoelina Minaya Ventura (Universidad Autónoma de Madrid)

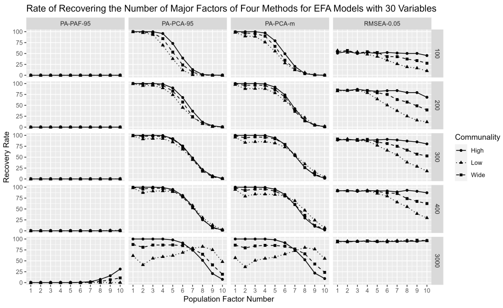
Psychological measurement is typically based on rating scales comprising a few ordered categories. When the mean of the observed responses approaches the upper or the lower bound of the scale, the distribution of the data becomes skewed and, the Pearson correlation between variables is attenuated, rendering an excessive number of factors. This poster presents the results of a simulation study investigating the problem of overfactoring and some solutions. We compare five widely known approaches: (1) The maximum-likelihood (ML) factor analysis model for normal data, (2) the categorical factor analysis (FAC) model based on polychoric correlations and ML estimation, (3) the FAC model estimated using a weighted least squares algorithm, (4) the mean corrected chi-square statistic by Satorra–Bentler to handle the lack of normality, and (5) the Samejima's graded response model (GRM) from item response theory (IRT). Likelihood-ratio chi-square, parallel analysis (PA), and categorical parallel analysis (CPA) are used as goodness-of-fit criteria to estimate the number of factors. Our results indicate that the ML estimation led to overfactoring in the presence of skewed variables both for the linear and categorical factor model, and that the Satorra–Bentler and GRM constitute the most reliable alternatives to estimate the number of factors.

Determining the Number of Factors in Exploratory Factor Analysis with Model Error

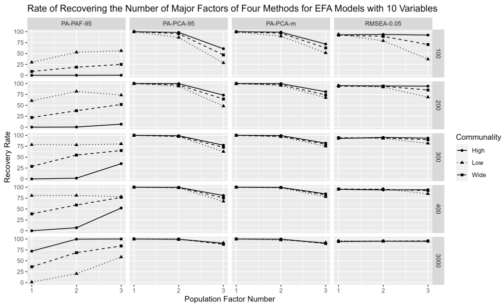
Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Yilin Li (University of Notre Dame), Prof. Guangjian Zhang (University of Notre Dame)

A key decision in exploratory factor analysis (EFA) is to determine the number of factors. Parallel Analysis (PA) and its variants are often recommended to aid this decision and their efficacy has been largely supported by simulation studies. The goal of the current study is to examine how PA and its variants perform in more realistic situations where EFA models fit approximately rather than perfectly. For comparison, we also consider a factor retention method that involves a model fit measure (Root Mean Square Error of Approximation, RMSEA) specifically designed to deal with model error. Our main findings include (1) PA is satisfactory when the factors are well-represented (high variable-to-factor ratios), but its performance becomes less satisfactory when the factors are not well-represented (low variable-to-factor ratios); (2) The RMSEA-based method is more satisfactory than PA under most conditions unless the sample size is very small; (3) The performance of the RMSEA-based method improves with larger samples, but the performance of PA and its variants do not improve with large samples.



B1range.upgrade555combine.pc.facet.j30.png



B1range.upgrade555combine.pc.facet.j10.png

Identifying predictors of careless responding trajectories over time

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Prof. Inés Tomás (University of Valencia), Dr. Ana Hernández (University of Valencia), Prof. Vicente González-Romá (University of Valencia), Ms. Clara Cuevas (University of Valencia), Prof. Anna Brown (University of Kent)

Careless responding (CR) is a critical issue that undermines data quality by reflecting insufficient attention to survey items (Podsakoff et al., 2012). Understanding whether CR constitutes a persistent characteristic or a momentary condition is essential for addressing CR prevention and management. Recent research (Tomás et al., 2024) showed varying CR patterns of trajectories over time, identifying four distinct subpopulations: two exhibiting consistent CR behavior (either careful or careless) and two showing temporal variations in CR (either increasing or decreasing). This study aims to explore the influence of sociodemographic factors (such as gender, age, and education) and personality traits (including agreeableness, conscientiousness, extraversion, neuroticism, and openness) on the likelihood of an individual's belonging to these CR subpopulations. Participants were 707 Spanish employees (50.4% men, aged 21 to 59). A within-subject longitudinal design with eight data collection points over 24 months, and multinomial logistic regression analysis were used. Results showed that age, education, and certain personality traits (agreeableness, extraversion, and neuroticism) significantly contributed to distinguishing between the two stable CR groups. Specifically, older individuals with higher education and agreeableness levels, and lower extraversion and neuroticism were more likely to be classified as careful respondents. Additionally, agreeableness and neuroticism were key in differentiating stable careless respondents from those with changing CR patterns over time. These insights highlight the complexity of CR and underscore the need for further research to uncover additional predictive factors, thereby enhancing our capacity to accurately identify and address CR behaviors.

This study was supported by the grant: PID2022-141339NB-I00/MCIU/AEI/10.13039/501100011033/FEDER,UE.

Estimation issues in growth curve models with short waves

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Yasuhiro Yamamoto (The Joint Graduate School (Ph.D. Program) in Science of School Education Hyogo University of Teacher Education), Dr. Yasuo Miyazaki (Virginia Polytechnic Institute and State University)

Growth curve model is a popular approach for studying change/growth over time in educational and psychological studies. In the growth curve model, the number of waves is typically short such as three, four, or at most ten waves. In research on multilevel models using cross-sectional data, it is known that extremely small average cluster size, such as two, could produce bias on variance component parameters even though the number of clusters is relatively large (e.g., Clarke, 2008; McNeish, 2014). It is not well known, however, whether unbiased estimates, especially for the variance component parameters, can be obtained or not in the growth curve model with short waves. Thus, in this study, we investigated this issue via Monte Carlo simulation. Data from linear growth curve model with correlated random intercepts and slopes with differing number of waves were generated, and the parameter recovery was examined. Analysis was conducted using Restricted maximum likelihood (REML) and Full maximum likelihood (FML), and we compared the results. In the preliminary results, it was indicated that the short waves, such as five waves, produced some bias on variance component parameters, specifically positive bias on the slope variance and negative bias on the covariance/correlation. Since, in the growth curve model, it is important to estimate variance component parameters such as slope variance and correlation between intercept and slope in order to understand the nature of the change over time and the individual differences, its implications for the practice were discussed.

Model-based Test and Item Analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Timo Bechger (TCS), Prof. Gunter Maris (TCS)

Test and item analysis involves summarizing data from a test administration to guide test construction. Traditional TIA is based on classical test theory and includes simple statistics, like item facilities, item-total correlations, etc., supplemented by graphics such as distractor plots. Although TIA is an important aspect of our work, little progress has been made since the 1950s. This talk is about the use of dedicated statistical models to make TIA more insightful and useful for test construction.

Utilizing the Bayesian Networks' structural learning algorithm to estimate Q-Matrix in Cognitive Diagnosis Models

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Lingling Wang (Shenyang Normal University), Prof. Tao Xin (Beijing Normal University), Prof. Bo-Nan Jiang (Shenyang University of Technology)

Cognitive diagnosis models (CDMs) refer to a set of psychometric models that intend to group individuals into latent classes with distinct skill profiles. A more generic term for skills is 'attributes', which are typically assumed to be binary latent variables. An attribute profile indicates which attributes individuals have possessed and which they have not. In CDM, a Q matrix associating each item in a test with the cognitive attributes is necessary to derive the attribute profiles of individuals. Defining the Q matrix is the most fundamental step in cognitive diagnosis. The conventional way of calibrating cognitive attributes mainly relies on the subjective judgments of experts. This research proposed a method using a structural learning algorithm in a Bayesian network (BN) to estimate the Q matrix. Both the viability and efficacy of the suggested approach were examined by running simulations and conducting analysis based on real data. The outcomes of the simulation indicated that the proposed approach mostly exceeded the performances of the available methods, and its advantage not only stemmed from better estimation accuracy but also its computational efficiency. The efficiency of the proposed approach was also verified by real data analysis. When compared to the stepwise method, the model data fit of the estimated Q matrix by the BN method in empirical data is more satisfactory and the RMSEA is lower. Consequently, this Q-matrix estimation method based on BN can improve the accuracy and computational efficiency and promote the application of cognitive diagnosis assessment in practice.

Estimating nonlinear effects of random slopes with Multilevel SEM

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Sarah Humberg (University of Münster), Simon Grund (University of Hamburg), Prof. Steffen Nestler (University of Münster)

In diverse psychological disciplines, researchers increasingly focus on person characteristics that relate to within-person associations (WPAs) of variables that fluctuate over time. For example, a person's stress reactivity can be operationalized as the association between repeated measures of their stress level and their momentary negative affect. Accordingly, researchers are interested in relating WPAs to potential person-level outcome variables (e.g. job performance). Such effects of WPAs can be tested with a multilevel structural equation model (MSEM), in which the WPAs are specified as random slopes on Level-1, and the latent representations of the slopes are used to predict the outcome variable at Level-2. In the present research, we consider the case of WPA-effects that are non-linear - for example, a U-shaped effect of one WPA or a moderation effect of two WPAs. The corresponding MSEM thus contains latent interaction terms of the random slopes, which complicates the model's estimation, and which may affect the quality of the parameter estimates and inference in yet unknown ways. I will report on a simulation study in which we evaluated the performance of the non-linear MSEM approach for different classes of non-linear effects (U-shaped; moderation), and for different sample sizes at both levels (individuals, time points). In addition, we compared the MSEM approach with three simpler approaches: a manifest 2-step approach, a 2-step approach using a single-indicator model at Level-2, and a plausible-values approach. Besides presenting the simulation results, I will derive recommendations for practice.

Differential item functioning: Effect sizes classification

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Michaela Cichrová (Institute of Computer Science, Czech Academy of Sciences and Faculty of Mathematics and Physics, Charles University), Dr. Adéla Hladká (Institute of Computer Science, Czech Academy of Sciences), Dr. Patrícia Martinková (Institute of Computer Science, Czech Academy of Sciences and Faculty of Education, Charles University)

An important aspect of subsequent analysis in multi-item measurements involves checking for Differential Item Functioning (DIF), that is, identifying potentially biased items that function differently across distinct population groups. Numerous statistical procedures have been developed to detect DIF by testing the statistical significance of the item-level differences between groups. However, besides the statistical significance of DIF, it is also vital to assess its practical significance by examining the magnitude of the corresponding effect size measure. This is necessary because even practically "negligible" differences can be statistically significant.

In this work, we review existing DIF effect size measures and the cut-off values used to classify the effect size magnitudes as "negligible" (A), "moderate" (B), and "large" (C) for logistic regression models and Item Response Theory models, both types of models in the case of binary items. The properties of the effect size measures and cut-off values are evaluated through a simulation study for both uniform and non-uniform DIF. Based on the simulation study, several effect size measures seem to display unsatisfactory properties (e.g., dependence on sample size, inconsistent classification of the underlying DIF, underestimating of the true underlying effect size measure). We propose solutions to observed inconsistencies and issues.

How to (not) successfully separate trait and response style parameters in IRTree models

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Viola Merhof (University of Mannheim)

It is a well-acknowledged concern in psychometrics that when responding to self-report Likert-type items, respondents do not always answer on the basis of trait of interest alone, but that response styles can influence their decisions. IRTree models aim to separate the influences of trait-based and response style-based response processes by decomposing the ordinal rating responses into sub-decisions that are assumed to be made based on either of such processes. Here we show that under certain conditions, the meaningful separation of trait and response style parameters by IRTree models may be impaired in that the response style factor mimics the trait and takes over part of the trait-induced variance in item responding. As a result, the substantive meanings of the estimated response style parameters may not correspond to the meanings attributed to them, that is, content-unrelated category preferences. Simulation analyses and an empirical example investigate the causes and consequences of such biases for the validity of interpretations drawn from the data.

Leveraging Factor Copula Models to address non-normality and heavy-tailed distributions in cyberbullying data

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Michaela Varejkova (Institute of Computer Science, Czech Academy of Sciences; Faculty of Mathematics and Physics, Charles University)

In recent years, the application of Factor Copula Models for item response data (Nikoloulopoulos & Joe, 2015) has been gaining attention in psychometrics, offering a robust alternative to traditional methods like IRT models. Factor Copula Models allow to relax normality assumptions inherent to IRT, making them particularly advantageous when dealing with data characterized by heavy-tailed distributions. Unlike IRT, which relies on Gaussian assumptions and may struggle with extreme observations, Factor Copula Models provide a flexible framework capable of capturing complex dependencies among items by accommodating non-Gaussian distributions. An accurate description of tail dependence is crucial for understanding rare but impactful events. In this study, we employ Factor Copula Models to explore the multidimensional structure of cyberbullying behavior using data from the Trends in International Mathematics and Science Study (TIMSS) 2019 questionnaire. Cyberbullying has emerged as a pervasive issue with significant psychological and social implications, yet its complex nature presents challenges for measurement and analysis, necessitating advanced statistical methods. By leveraging Factor Copula Models, we demonstrate improved model performance and enhanced robustness to extreme responses often encountered in cyberbullying research. Through a comparative analysis with IRT, we highlight the advantages in accurately modeling the underlying structure of cyberbullying data, especially in instances where traditional methods fail due to non-normality and heavy-tailedness. We also investigate the influence of demographic and socioeconomic factors on the dependence structure of cyberbullying behaviors, shedding light on potential risk profiles and moderators.

A new approach for detecting prosopagnosia in children

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Prof. Alexander Avian (Medical University of Graz)

Prosopagnosia is a disorder characterized by severe difficulty recognizing faces. There are currently two main approaches used to diagnose prosopagnosia: self-report and testing. During the tests, different faces are given that have to be recognized. These faces either come from famous people or had to be learned beforehand. Learning faces requires a certain cognitive development and recognizing famous faces requires that these people are actually known. Both represent a challenge, especially for younger children. This study examines the ability of family member faces (FMF) in contrast to altered family member faces (aFMF), and stranger faces (SF) to identify individuals with prosopagnosia. A set of 143 faces (FMF: n=62; aFMF: n=54; SF: n=27) was specially compiled for a family with several people with prosopagnosia. Seven members from this family (3 prosopagnosia) evaluated these faces. Overall 88% of the FMF, 85% of the SF but only 56% of the aFMF were correctly identified with a worse performance of family members affected by prosopagnosia (80%, 70%, 40% vs. 91%,91%, 63%). People with prosopagnosia often confused family members (up to 24% vs. up to 3% of FMF) or mistook aFMF or SF for family members (up to 63% vs. up to 26%). Furthermore, they did not recognize the correct sex in up to 12% of cases. This new approach appears to be a promising strategy for identifying people with prosopagnosia. Methodological challenges such as standardization and definition of suitable psychometric properties based on the images still need to be solved.

Interpretation of change scores in latent change score models

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Una Mikac (University of Zagreb Faculty of Humanities and Social Sciences)

The latent change score models are becoming more popular because of the multiple information they provide about the longitudinal data (e.g., Matusik et al., 2021). The central component of these models is the latent change score (LCS) that represents the change between two consecutive time points defined on a latent level. One of the most important contribution is the possibility to regress LCS on predictors in the previous time points. However, interpretation of these regressions can be problematic and should take into account the direction of change. If the outcome variable has for most individuals increased from one time point to the following, as indicated by a positive LCS, a positive regression parameter indicates that the predictor predicts an increase in the outcome variable. However, if the average LCS is 0 and variance significant, then for some individuals there is an increase, and for others a decrease, making it unclear what the regression parameter indicates. Does the predictor predict an increase regardless of the direction of change, or does it predict a larger change for those with positive LCS and a smaller change for those with negative LCS? These issues will be explored on data from a three-wave study (*N* = 1011, 56.8% female; mean age 41.4) of temporal precedence of burnout dimensions measured by Burnout Assessment Tool (Schaufeli et al., 2020). We suggest a graphic representation based on estimated means and variances of LCS and their predictors as a way to help interpret the meaning of estimated regression coefficients.

Exploring youth mental health with IRT: proposal of GHQ-12 reassessment

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Anna Comotti (IRCCS Foundation Ca' Granda Ospedale Maggiore Policlinico), Teresa Barnini (IRCCS Foundation Ca' Granda Ospedale Maggiore Policlinico), Alice Fattori (University of Milan), Maria Emilia Paladino (University of Milan-Bicocca), Michele Augusto Riva (University of Milan-Bicocca), Matteo Bonzini (University of Milan), Micheal Belingheri (University of Milan-Bicocca)

The General Health Questionnaire-12 (GHQ-12) is a widely used screening tool for mental health assessment however its traditional scoring methods and cutoffs may not adequately capture the mental health complexities of younger populations.

This study explores GHQ-12 responses from a sample of university students. Possible differences in means scores considering gender, age, academic field and degree course were assessed through t-test or one-way ANOVA as appropriate. To deeper understanding different levels of severity and individual item impact on general distress measurement, we applied Item-Response-Theory (IRT) techniques, including the 2-PL model and the LC (Latent Class)-IRT model.

A total of 3834 university students participated in the study. Results showed that a significant proportion (79%) of students reported psychological distress. Females and younger students obtained significantly higher average scores compared to others. IRT analysis found item-specific variations in mental distress levels, with more indicative items for short-term fluctuations and potential severe mental health concerns. Latent class analysis identified three distinct subgroups among students (including 20%, 37%, 43% of the participants respectively) with different levels of psychological distress severity. Comparisons with a population of 990 healthcare workers, whose psychological distress was measured through the same questionnaire (frequently used to monitor the mental health status of workers in different occupational setting), highlighted the unique mental health challenges faced by students.

We suggested a reevaluation of GHQ-12 applicability and cutoff scores for younger populations, emphasizing the need for accurate instruments in mental health evaluation.

Mapping methodological variations in ESM research: A systematic review

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Lisa Peeters (KU Leuven), Dr. Richard Artner (KU Leuven), Prof. Wim Van den Noortgate (KU Leuven), Prof. Ginette Lafit (KU Leuven)

In recent years, the Experience Sampling Method (ESM) has become a widespread tool to study time-varying constructs (e.g., emotions, substance craving) across many subfields of psychological research (e.g., organizational psychology, clinical psychology). This large variety in subfields of research and constructs of interest has led to considerable methodological variation. This issue of methodological variation has been acknowledged by many researchers, but few have attempted to systematically assess the methodological choices made by ESM researchers in psychology, and to question the motives behind these choices. Existing systematic reviews have focused on specific demographics (e.g., certain clinical populations) or subfields of psychology, while previous non-systematic attempts to map methodological variation are limited, outdated and focused exclusively on (pre-data-collection) design choices. Therefore, the first aim of the current systematic review is to describe the methodological variation (from conception of the research question to data analysis) in ESM designs in the recent psychological literature. The second aim of the review is to assess the quality of ESM studies, which encompasses complete and transparent reporting, open science practices, as well as the validity of the methodological choices made by ESM researchers (e.g., the (mis)match between the research question and the type of statistical analysis). This systematic review – and its aim to *describe* the content and quality of ESM research – is a first step towards a broader goal to *improve* the methodological quality of ESM research in psychology.

Multi-method evaluation of the predictive utility of self-report measures

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Hannah Lewis (University of North Carolina at Chapel Hill), Dr. Halpin, Peter Francis (University of North Carolina at Chapel Hill), Dr. Bryant Hutson (University of North Carolina at Chapel Hill)

Recent initiatives to broaden participation in science, technology, engineering, and mathematics (STEM) education have led to the widespread use of self-report measures related to STEM persistence (e.g., identity, self-efficacy). However, it is unclear whether self-report measures of STEM persistence do indeed predict more distal outcomes related to retention in STEM fields (e.g., graduation with a STEM degree). The current study seeks to compare different methods for modeling the predictive validity of one widely used self-report measure of STEM persistence: the Persistence in the Sciences (PITS) survey. We compare logistic regression, random forest, and deep learning techniques to predict whether students graduated with a STEM degree. We focus on evaluating the extent to which the different modeling approaches provide empirical evidence about the predictive validity of the PITS survey in a complex administrative dataset. The sample includes *N = 1574* students who took the PITS survey at two assessment time points (pre- and post-semester for participating courses) over the years 2017-2022. The dataset also includes many potential control variables (course grades, major), moderators (student demographics), and unobserved sources of heterogeneity (e.g., cohort effects during the COVID-19 pandemic), which provides an interesting and important real-world context for comparing the methods.

Quantifying Predictive Uncertainty in Validity Studies

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Youmin Hong (University of Maryland), Youngjin Han (University of Maryland), Dr. Yang Liu (University of Maryland), Youjin Sung (University of Maryland), Dr. Ji Seung Yang (University of Maryland)

It is important to gather empirical evidence of validity in practices of scale development. Predictive validity is a core component of validity: Scores obtained from measurement instruments are expected to predict criterion variables that are in theory pertinent to the constructs to be measured. It has been argued recently that raw item scores, when used in conjunction with modern regression techniques (e.g., regularization and regression trees), exhibit higher predictive utility (measured by, e.g., cross-validation error) than aggregated scale scores (e.g., summed scores and factor scores) or latent variables in measurement models (e.g., factor analysis and item response theory). We, however, find this argument incomplete, lacking quantification of predictive uncertainty. In psychological and educational testing, uncertainty arises from not only imperfect measurement but also sampling variability and model misspecification. It is then crucial to evaluate all predictive models with a universal inferential tool that simultaneously characterizes various sources of uncertainty. In the current work, we reassess the predictive utility of item scores, scale scores, and latent variables from the perspective of conformal prediction (CP), which is a model-free uncertainty quantification framework that yields predictive inference with finite-sample guarantees. We conduct a Monte Carlo experiment to evaluate and contrast CP sets resulted from different types of scores and different predictive regressions under various conditions.

Performance of parallel analysis in bifactor model with ordinal items

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Hyunjung Lee (Fordham University), Dr. Heining Cham (Fordham University)

Bifactor models, latent structural models that include a single general factor and specific factors, are commonly utilized in behavioral research, especially personality and psychopathology (Reise, 2012; Rodriguez et al., 2016). However, bifactor models tend to exhibit a better fit than the common factor models (Luciano et al., 2020), so exploring other criteria for assessing dimensionality is needed. Many studies assessed the performance of parallel analysis (PA) for common factor models (e.g., Lim & Jahng, 2019) or bifactor models with continuous data (e.g., Jiménez et al., 2023). However, studies on the dimensionality of bifactor models with ordinal data are limited. Thus, this study will explore the performance of PA on bifactor models with ordinal data using Monte Carlo simulation by comparing the results with continuous data. For this purpose, population models varying in the number of specific factors (3 and 5), the number of indicators per specific factor (4 and 6), the magnitude of general factor loadings (0.3, 0.5, and 0.7), the loading between the general factor and specific factors (0.3, 0.5, and 0.7), the sample size (200, 500 and 1000), and the number of response category (2, 5, and continuous) will be examined. For each of the 324 conditions, 1000 replications will be generated. The R software (R core team, 2023) will be used for data generation and analysis, and the performance of PA of both continuous and ordinal items will be reported. This study aims to provide valuable insights for researchers seeking to employ PA for bifactor models.

Comparing traditional measurement invariance and alignment approaches to data harmonization

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Meltem Ozcan (University of Southern California), Dr. Hok Chio (Mark) Lai (University of Southern California)

More data are available to and accessible by researchers than ever before thanks to open science initiatives coupled with recent advances in technology that have enabled the convenient storage and widespread distribution of large files. While this development has opened the path for innovative lines of inquiry involving the simultaneous analysis of multiple datasets, a number of measurement issues need to be resolved before inferences can be deemed valid, such as noninvariance and unreliability. A number of methods have been proposed for data harmonization with psychological measures. Here, we compare traditional measurement invariance vs. alignment optimization approaches in harmonizing math self-efficacy scores across two national data sets in the U.S.: the Education Longitudinal Study of 2002 (ELS) and the High School Longitudinal Study of 2009 (HSLS). Math self-efficacy is measured with 5 items in ELS, but only 4 items in HSLS with small differences in item wordings. In the traditional approach, likelihood ratio tests found all items to be noninvariant, and we chose a best-fitting partial scalar invariance model for harmonization. On the other hand, the alignment approach does not assume exact invariance of any item. For both approaches, we computed Bartlett factor scores, and documented differences between the two approaches in terms of score distributions, reliability, and correlations with other variables. Our analyses provide insights on the use of invariance analyses for data harmonization when items are non-overlapping across samples, and methodology for reliability estimation with harmonized scores.

Test Length Optimization with Deep Reinforcement Learning

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. James Zoucha (University of Northern Colorado), Dr. Igor Himelfarb (National Board of Chiropractic Examineers), Dr. Nai-En Tang (National Board of Chiropractic Examineers)

This study explores the use of a Deep Reinforcement Learning (DRL) algorithm to determine the optimal number of items needed for the National Board of Chiropractic Examiners (NBCE) Part I exam. The goal was to find, or confirm, the number of items needed to accurately and consistently estimate student ability under established content and exposure constraints. The DRL algorithm was modeled to resemble a computer-based test and trained on 100,000 episodes. Despite achieving some success, none of the 10,000 additional episodes where the trained algorithm was used to find subsets of items comprising a possible test met all the desired specifications. The study recommends maintaining the original test length of 240 items until further adjustments can demonstrate that a test with fewer items can maintain domain and difficulty constraints. Future research may benefit from providing a larger item bank, obtaining more computational resources and exploring different algorithm architectural choices.

Robustness of balanced item parceling in treating acquiescence in SEM

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Jun-Ting Liu (National Taiwan University), Mr. You-Lin Chen (National Taiwan University), Dr. Li-Jen Weng (National Taiwan University)

Acquiescence (ACQ), characterized by a consistent extreme response tendency regardless of item content, often occurs in Likert-type scales. In structural equation modeling (SEM), ignoring ACQ can distort covariances among variables, resulting in biased parameter estimates and poor model fit (Savalei & Falk, 2014). Balanced scales containing an equal number of positively and negatively worded items are often used to control for ACQ. Based on balanced scales, Weijters and Baumgartner (2022) recently proposed using balanced item parcels (BIP) to deal with ACQ in SEM. BIP averages over an equal number of positively and negatively worded items measuring the same construct and assumes that items are equally affected by ACQ. This equal impact assumption has been challenged in the literature (Ferrando et al., 2003). In this study simulation was conducted to assess the robustness of BIP approach in violation of the equal impact assumption. ACQ is represented by a method factor affecting all the items and the equal impact assumption is satisfied when the method loadings are equal. We used a five-factor full structural equation model to generate data under combinations of equality of method loadings (equal, mildly unequal, moderately unequal), sample size (300, 400, 500) and number of items (32, 64). The biases and standard errors associated with structural parameter estimates suggested that the BIP approach to treat ACQ in SEM appeared to be robust against the violation of equal impact assumption.

Testing minor factors incremental value in an essentially unidimensional measure of work-family enrichment

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Prof. Pieter Schaap (University of Pretoria), Prof. Eileen Koekemoer (University of Pretoria), Prof. Marissa Brouwers (North West University)

Introduction

Work-family enrichment constructs typically considered as multidimensional lack empirical support. However, research suggests quality-of-life measures are mostly essentially unidimensional, consisting of an underlying general factor that inherently tap into multiple minor domains or factors. Recent studies indicate that Brouwer's MACE work-family enrichment scale aligns with this unidimensional perspective. However, conventional statistical methods tend to underestimate the importance of these minor factors.

Aim of study

This study aimed to examine whether the MACE constructs are essentially unidimensional, with incremental and distinct value in minor factors, using contemporary statistical indicators.

Sample and statistical analyses

The MACE measure was applied to 627 South African employees across diverse industries. Bayesian bifactor SEM (Bi-BSEM), Omega Hierarchical (HS) indices, explained common variance of the specific factor (ECVss) indices, Dueber's (2023) new criteria and the Unival package in FACTOR programme were employed to assess the essentially unidimensional model and the incremental and distinct value of sub-factors in relation to the outcome variables work vigor, work dedication, career satisfaction, and job satisfaction.

Results

The results indicated that the MACE is essentially unidimensional with sub-factors, specifically work-family perspective and work-family effect, having a significant incremental and distinct relation to work vigor and work dedication.

Conclusion

This study established that the MACE can be regarded as essentially unidimensional, with two sub-factors offering meaningful interpretations and added value to specific outcomes using contemporary statistical indicators and methods. This approach facilitates the simultaneous examination of work-family enrichment as a comprehensive theoretical concept alongside its minor components, enriching theory development.

Exploring the Feasibility of Automatic Item Generation for Korean Language Assessment Considering Item Types

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Haneul Lee (Inha University), Dr. Yongsang Lee (Inha University)

Recently, there is a growing attention on automatic item generation using generative artificial intelligence in the field of education to alleviate the time and cost burden associated with test item development. Considering the necessity to create passages along with test items in language assessment such as English and Korean, there is a need to research the potential of using artificial intelligence for automatic item generation in language assessment. Accordingly, this study aims to explore the possibility of automatic item generation in Korean language assessment in various item types using ChatGPT and to exam the item quality generated by ChatGPT. For this purpose, we selected item types from the National Assessment of Educational Achievement (NAEA) in South Korea and utilized ChatGPT to generate items corresponding to each item type based on example data of passages and items. Furthermore, we compared the cosine similarity between items generated by ChatGPT and the original items. Also, we evaluated the quality of ChatGPT-generated items through expert review. We also discussed the potential and limitations of using ChatGPT for automatic item generation in Korean language assessment.

Item and scale invariance: Comparing at-risk post-secondary to school students

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

John Sabatini (The University of Memphis), Ryan Kopatich (Northern Illinois University), John Hollander (The University of Memphis), Daniel Feller (The University of Memphis)

Introduction

In this study, we administered a battery of reading skills subtests to college students identified as at-risk of academic failure. The battery was designed for use with grade 5-10 students falling behind grade level expectations, with vertical-IRT-based scales for each subtest. In this presentation, we examine the degree to which the items and scales can appropriately be used with a college population, versus developing new items or scales.

Methods

Vertical IRT-based (2pl) scales were constructed for each of 6-subtests from a sample of ~31,000 US grades 5-10 students. Reliabilities for each subtest range from .65-.96.

College student sample: we tested multiple waves of at-risk freshman across multiple postsecondary settings (n=~2300) in 2022-24.

Analysis: We first use the IRT-2PL fixed parameters of the secondary sample to conduct a calibration analysis of the post-secondary sample. We are currently also creating new item parameters and scale scores based on the college sample only. We will show results comparing the two populations including item fit statistics, scale properties, etc.

Results: Preliminary results are complex. About 20% of the college student sample score at or near ceiling levels on the difficult subtest forms, suggesting we will need to create more difficult items. There is significant variability for most students across subtests. In the session, we will present more detailed results for each subtest.

Conclusions: This research addresses issues of generalizability and validity when adapting reading skills tests designed for school students for use with post-secondary at-risk students.

Psychosocial Factors Influence Undergraduates' Mental Health: A SEM Mediation Analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Yi-Jou Chen (National Taiwan University), Dr. Grace Yao (National Taiwan University)

Background: In recent years, self-harm behaviors among undergraduate students have been reported more frequently, indicating the urgent need to address undergraduate students' mental health issues. This research investigates the mechanism of social support as a social factor and personal traits (self-esteem and psychological resilience), regulatory emotional self-efficacy (RESE) as psychological factors on National Taiwan University (NTU) undergraduate students' mental health (negative emotional responses, well-being and life satisfaction). In this research, we assessed the mediation effect of RESE on the relationship between social support and mental health and the serial mediation effect of personal traits and RESE on this relationship by 2 SEM mediation models. **Methods:** In total, 3164 students participated in this study. We used reliable and validated instruments to measure social support, self-esteem, psychological resilience, RESE, negative emotional responses, well-being and life satisfaction. SEM with bootstrapping was used to examine proposed mediation effects. **Results:** The goodness of fit of the 2 SEM mediation models was acceptable. Mediation analysis indicated that RESE negatively mediated the relationship between social support and negative emotional responses (β= -.46, $p < .001$); RESE positively mediated the relationship between social support and well-being (β= .40, $p < .001$) and life satisfaction (β = .43, $p < .001$). In addition, personal traits (self-esteem and psychological resilience) and RESE mediated the relationships above serially. **Conclusion:** The research results contribute to an understanding of the effects of psychosocial factors on the mental health of undergraduate students, serving as a reference for schools and practitioners in counseling and intervention.

Evaluation of GAIN's performance in handling missing data in SEM

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Luqi He (McGill University), Bellete Lu (McGill University), Dr. Carl Falk (McGill University), Dr. Heungsun Hwang (McGill University)

Structural equation modeling (SEM) is widely used to specify and examine the relationships between observed and latent variables based on theory. In practice, it is often inevitable to encounter missing data for various reasons, including nonresponse, attrition, etc. In the presence of missing data, SEM can lead to biased estimates and loss of efficiency. Full information maximum likelihood (FIML) and multiple imputation (MI) are commonly used to deal with missing data in SEM. However, their performance has scarcely been examined when the proportion of missing data is high (e.g., > 50%). Generative adversarial imputation nets (GAIN) have recently been proposed to impute missing data in machine learning. It has shown promising results, particularly when the proportion of missing data is high. However, its performance in SEM remains unexplored. To fill this gap, we conducted a simulation study to systematically evaluate GAIN's capability to handle missing data in SEM. We manipulated various experimental factors, such as sample size, model misspecification, and the proportion of missing data, and compared GAIN's performance in parameter recovery with FIML and MI.

A Two-Step Q-matrix Estimation Method

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Hyunjoo Kim (University of Illinois, Urbana-Champaign), Prof. Hans Friedrich Koehn (University of Illinois, Urbana-Champaign), Prof. Chia-Yi Chiu (Teachers College, Columbia University)

Cognitive Diagnosis Models in educational measurement are restricted latent class models that describe ability in a knowledge domain as a composite of latent skills an examinee may have mastered or failed. Different combinations of skills define distinct latent proficiency classes to which examinees are assigned based on test performance. Items of cognitively diagnostic assessments are characterized by skill profiles specifying which skills are required for a correct item response. The item-skill profiles of a test form its Q-matrix. The validity of cognitive diagnosis depends crucially on the correct specification of the Q-matrix. Typically, Q-matrices are determined by curricular experts.

However, expert judgment is fallible. Data-driven estimation methods have been developed with the promise of greater accuracy in identifying the Q-matrix of a test. Yet, many of the extant methods encounter computational feasibility issues either in the form of excessive amounts of CPU times or inadmissible estimates.

This presentation introduces the Two-Step Q-matrix Estimation (TSQE) method for estimating the Q-matrix of cognitive diagnostic assessments from item responses that may conform to any cognitive diagnosis model. The TSQE method combines a provisional attribute extraction (PAE) algorithm to build a provisional Q-matrix from the data at hand and a refinement-and-validation algorithm for optimizing the provisional Q-matrix. Results of large-scale simulation studies for evaluating the performance of TSQE under multiple experimental conditions show that the new method outperforms extant estimation algorithms in accuracy and computational efficiency.

A visualization method to describe test equating procedure

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Haruhiko Mitsunaga (Nagoya University)

When we construct an item bank which contains item parameters assuming item response theory (IRT), the procedure of test equating should be thoroughly documented. However, sometimes the document is too complicated if the flow of procedure is written in text. For example, when we use multiple test forms to obtain item parameters on the common scale, a common-item nonequivalent groups design is usually adopted, which means that the flow of computation process should be illustrated in an intuitive way without any ambiguity. In this presentation, a new protocol to depict the procedure of obtaining a common scale by drawing a diagram with several icons is proposed. The diagram can describe not only the procedure of test equating, but also a list of concordance between the actual procedure and a pseudocode of computer program for estimating item parameters and linking constants. The proposed method can also illustrate the dependency among the process of analyzing datasets, which is useful for searching for a critical path. Practical examples for describing the procedure of obtaining a common scale are illustrated as a summary of drawing actual test design, such as common-item design, nonequivalent group design and anchor test design.

Existence and uniqueness of MLE of the ability in IRT

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. CHE CHENG (National Taiwan University), Mr. Hau-Hung Yang (National Taiwan University), Prof. Yung-Fong Hsu (National Taiwan University)

This research investigates the conditions for the uniqueness and existence of maximum likelihood estimator (MLE) of the latent trait in item response theory (IRT). The popular IRT models, including the one-parameter and two-parameter logistic and normal ogive models, are special cases. We prove that, assuming log-concavity for the family of item response functions,

participants' responses being not all-correct or all-wrong is necessary and sufficient for the existence and uniqueness of MLE. Also, we note that while the MLE can theoretically exist and be unique, the commonly used Newton-Raphson algorithm might not always converge. To overcome this hurdle, we incorporate the bracketing method into a newly developed algorithm, called IRTMLE, to ensure the finding of MLE in various cases. IRTMLE also can handle the scenario where participants' responses are uniformly zeros or ones.

Uncovering efficient reasoning strategies in interactive inquiry tasks using sequence mining

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Shuang Wang (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University), Dr. An Hu (State Key Laboratory for Artificial Microstructure and Mesoscopic Physics, School of Physics, Peking University), Dr. Wei Tian (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University), Dr. Tao Xin (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University)

Scientific inquiry is fundamental to science education as it fosters critical skills such as scientific reasoning and argumentation. However, the specific strategies that contribute to successful inquiry have not been thoroughly investigated. This study examines the relationship between students' scientific reasoning behaviors and their problem-solving efficiency within interactive scientific inquiry tasks. We analyzed the behaviors of 86 fourth-grade students who successfully solved a scientific problem in the interactive task. Using k-means clustering, the students were categorized into two groups, efficient and less efficient, based on their task completion efficiency. The reasoning behaviors were meticulously captured in log files and were subsequently recoded to reflect their task relevance. A set of comparative analytical techniques, including independent t-tests, sequence visualization, and sequential pattern mining, was applied to discern strategic differences between the two groups. The results showed that the less efficient group gathered more irrelevant evidence, conducted more experimental trials, and formulated more incorrect hypotheses compared to their more efficient counterparts. Furthermore, the less efficient group exhibited behavioral patterns suggesting a lack of strategic planning. These results highlight the role of metacognitive skills in successful and efficient scientific inquiry and suggest the need for tailored support to facilitate the acquisition of scientific reasoning skills.

Detecting DIF in PISA Reading Frequency Measurement

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Qing Cai (University of California, Berkeley), Mr. Mingfeng Xue (University of California, Berkeley)

Large-scale international assessments rely on invariance measurement invariance across genders for validity and fairness. This study scrutinizes the internal properties and equity of the Reading Frequency measurement in the PISA 2018 US Student Questionnaire, a critical area deserving more attention. PISA's questionnaire aims to gauge latent constructs, including reading frequency across five text types (magazines, non-fiction, fiction, newspapers, and comics). This approach was designed to furnish stakeholders with valuable insights into the interplay between reading habits and academic performance, thereby enriching the interpretation of assessment outcomes.

Previous research highlights gender disparities in reading preferences (Thums 2020), thereby casting a critical spotlight on the measurement design employed by PISA. Our study identified intermediate Differential Item Functioning (DIF) (.566) for one out of five items, thereby warranting a thorough reevaluation of reading frequency measurement practices on a broader scale.

In response to our findings, we advocate for strategic interventions aimed at mitigating DIF effects and enhancing the overall integrity of reading frequency measurement. Such measures may entail the recalibration of questionnaire items to ensure balance across genders or the judicious removal of items exhibiting DIF. By adopting a proactive stance in addressing measurement challenges, we can fortify the validity and fairness of large-scale assessments, ultimately advancing the cause of educational equity and excellence.

Predictive metrics in multilevel models with continuous outcomes

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Diego Iglesias (Universidad Autónoma de Madrid), Dr. Miguel A. Sorrel (Universidad Autónoma de Madrid), Dr. Ricardo Olmos (Universidad Autónoma de Madrid)

Multilevel Models (MLMs) have become a valuable research tool in the field of psychology due to their ability to effectively analyze nested data structures commonly found in the behavioral sciences, such as students nested within schools. MLMs enable a nuanced examination of individual and cluster-level predictors with respect to an outcome of interest. When working with continuous outcomes, measures of explained variance, such as R^2, provide a summary of the magnitude of the effects analyzed. Nevertheless, unlike single-level regression analysis, MLMs R^2 become more intricate when considering sources of variance at different levels. Recently, Rights & Sterba (2019) developed an integrative framework of R^2 measures in MLMs. Based on a completely full partitioning of variance, this framework offers a unifying approach to interpreting and choosing among the available R^2 measures considering specific research questions. However, the finite sampling properties of these R^2 measures across different conditions have not yet been exhaustively studied. The present study assesses the performance of the different R^2 measures as estimators of their respective population values through a Monte Carlo simulation. Among other factors, we examined the impact of sample size and the number of predictors at different levels. As expected, results suggest that the accuracy of the estimation diminishes as the model becomes complex and the sample size limited. We offer guidelines on minimum requirements for accurate estimation and discuss predicting unknown observations beyond the sample where the model was fitted, proposing the framework's application for out-of-sample predicted variance in MLMs.

The performance of latent mean estimates: Comparing full and partial scalar invariance

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Cheng-Hsien Li (National Sun Yat-sen University)

The establishment of measurement invariance is recognized as an important pre-condition for cross-group mean comparisons. The full development of multiple-group confirmatory factor analysis has greatly facilitated the practice of measurement invariance testing over the past two decades. However, full invariance is statistically difficult to achieve and rarely appears in applications. A direct research question is the impact of using fully or partially invariant models when the comparison of latent means is the primary research concern. To date, little research has been done to inform best practices of making latent mean comparisons when varying degrees of noninvariance in measurement parameters are considered. This study aims to reveal the consequences of incorrectly full scalar invariance and the performance of truly partial scalar invariance with different numbers of referent variables on latent mean estimates under varying degrees of noninvariance in loadings and intercepts. A Monte Carlo simulation design was proposed, and the simulation factors included the percentage of noninvariant variables, the magnitude of noninvariance, the magnitude of latent mean differences, and sample size in a one-dimensional measurement model with ten continuous observed variables. Preliminary results indicated the bias of latent mean estimates decreased as the number of referent variables used in the anchor set increased. However, the performance of latent mean estimates was stable and acceptable when only a single referent variable was used in the anchor set (i.e., the bias of latent mean estimates remained within .03). As the cross-group latent mean difference increased, so did the bias.

Added value of subscale change scores for evaluating individual change

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Angelina Kuchina (Tilburg University), Dr. Wilco Emons (Tilburg University), Dr. Maria Bolsinova (Tilburg University)

Measuring individual change is one of the important goals in psychological and educational assessment. Individual change can be evaluated using change scores - the differences between test scores at two (or more) measurement occasions. These scores are commonly used for measuring change only for a single domain, although stakeholders are often interested in evaluating change for many related domains at the same time. It has been shown that using a separate scores on such domains (i.e., subscale scores) does not always provide more information than using a composite score which combines the highly related subdomains. In particular, subscores should be distinct and reliable enough to have added value over total scores. This question of subscales' added value is even more relevant in the context of change scores – since researchers have concerns about the reliability of these scores and the appropriateness of using them for measuring change. In addition, change subscores should be investigated separately since their properties might differ substantively from the properties of subscores at a single measurement occasion. In this presentation, we will discuss under which conditions using subscale change scores has added value for measuring change on a subdomain. Furthermore, at the individual level we will look at the added value of subscale scores for detecting individual change. Results from the simulation study will be presented. Based on these results we will provide guidelines for the practitioners for the use of subscales while evaluating change.

Refinement of Automatic Item Generation for Mathematical Questions: Applying Item Response Theory to Large Language Models

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Haruki Oka (The University of Tokyo), Prof. Tsunenori Ishioka (National Center for University Entrance Examinations), Dr. Kensuke Okada (The University of Tokyo)

Automatic item generation is a technology that uses natural language processing to generate questions about a domain of interest automatically. In recent years, large language models (LLM) have enabled the automatic generation of questions following the intended content and format. However, LLMs are unable to quantitatively and accurately represent design-critical item properties, such as difficulty and discrimination, using linguistic information. Therefore, in this study, we first fine-tuned LLMs using a) question-containing sentences and b) corresponding pre-estimated item parameters, and then generated the items. Here, the item parameters were estimated and equated by an item response theory model in advance. We conducted an empirical study for this theory using a multiple-choice test in Mathematics for K-12 6th grade students. We fine-tuned the LLM using the item-specific textual information and their parameter values. Then, we prompted the LLM for new question items that have specific target item parameter values. For validation of the proposed method, we evaluated the performance by collecting the examinee answer data and correlating the estimated and originally specified item parameters. We found that the proposed approach produced mathematical questions whose answers accurately reflected the nature of the items after we specified the target population of examinees. These observations suggest that fine-tuning LLM on the basis of the textual information of the items and their parameter values enabled appropriate automatic item generation.

Do visual analogue scales perform better than Likert-type scales?

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Eva Šragová (Masaryk University), David Elek (Masaryk University), Dr. Hynek Cígler (Masaryk University), Gabriela Kalistová (Masaryk University)

When producing self-report measure items, researchers can choose from many response formats. Besides the most common Likert-type response scale (LS), visual analogue scale (VAS) may present a good alternative. In this presentation, we empirically compare the performance of the two response formats using a within-subject counterbalanced experimental design. Two unidimensional measures were used (height and autonomy), in two versions (one using LS and one VAS). In three sessions, participants (N = 1003 young adults, mean age 24.4 years, 77% female) completed each of the four measures in different orders and with two different time intervals. Based on latent-variable measurement models, we will discuss the effects of two response formats on the dimensionality of the measures, their invariance, criterion validity, and several types of reliability (dependence, stability over time, and internal consistency). We also comment on other practical factors associated with the choice between LS and VAS (such as completion time, number of corrections, missing data, participant perceived accuracy, difficulty, and general preference).

Framework choice: Discrete (CDM) or continuous (MIRT) modeling in Psychometrics

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Miguel A. Sorrel (Universidad Autónoma de Madrid), Nerea Cano (Universidad Autónoma de Madrid), Scarlett Escudero (Universidad Autónoma de Madrid), Dr. Pablo Nájera (Universidad Pontificia Comillas), Dr. Rodrigo Schames Kreitchmann (National University of Distance Education), Dr. Francisco J. Abad (Universidad Autónoma de Madrid)

In empirical applications to date, researchers have typically opted for either a discrete modeling framework, cognitive diagnosis modeling (CDM), or a continuous one, multidimensional item response theory (MIRT), without directly comparing the two. While some prior research has explored the impact of attribute continuity on parameter recovery in CDM, there is a lack of studies that clearly document any potential loss, if it exists, when using CDM information for examinee ranking or MIRT information for discrete profile classification. This constitutes the primary focus of the current study, which employs Monte Carlo simulation to address this issue, manipulating factors such as sample size, number of assessed dimensions, and item quality, among others. Furthermore, we aim to evaluate the performance of measures capable of discerning, from the data, whether one modeling framework is more appropriate than the other, as exemplified by relative fit statistics. The dependent variables include measures of individuals' parameter recovery (e.g., % of correct classifications) and the selection rate for each model (CDM or MIRT) of various relative fit statistics (e.g., AIC, BIC), as well as measures specifically aimed at determining the degree of 'continuity' of the latent trait. Through this study, we endeavor to establish guidelines that illustrate the potential advantages of each approach depending on the assessment goal (ranking or classification) and the assessment characteristics. The findings will contribute to enhancing the understanding of the suitability of CDM and MIRT in various assessment contexts, thereby informing decision-making processes in educational and psychological research.

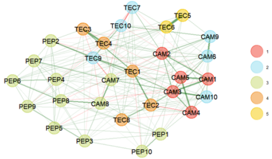
Examining the Complex Structure of OSCEs using Exploratory Graph Analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Nai-En Tang (National Board of Chiropractic Examiners), Dr. Igor Himelfarb (National Board of Chiropractic Examiners), James Zoucha (University of Northern Colorado)

Objective structured clinical examinations (OSCEs) are widely utilized in the health professions for their ability to assess various knowledge and competencies essential for proficient clinical practice (Sim et al., 2015). They are recognized for having complex structures as multidimensional practical exams (Petrusa et al., 1990). When scoring exams with complex structures, understanding the exam's dimensionality and selecting the appropriate scoring method are crucial prerequisites (De Ayala, 2009). Traditional techniques, such as factor analysis with parallel analysis or the Kaiser-Guttman eigenvalue rule, sometimes struggle to accurately estimate the number of dimensions, especially for exams with complex structures, such as OSCEs (Ruscio & Roche, 2012; Crawford et al., 2010). Consequently, this study aims to explore the structure of the chiropractic OSCE exam using a novel approach, exploratory graph analysis (EGA, Golino & Epskamp, 2017).

This study analyzed a single administration of the chiropractic OSCE exam involving 798 examinees. The exam comprised 30 stations and was built around three content domains: case management, post-encounter probe, and chiropractic technique. To address the risk of overfitting and enhance result generalizability, an initial EGA was followed by a 500 bootstrap EGA (bootEGA; Christensen & Golino, 2021) to assess structural stability across bootstrapped EGA outcomes. Variability in the number of dimensions and low item stability across replicate bootstrap samples was observed. These findings may suggest the presence of a complex structure and within-item multidimensional properties within the OSCE exam. The discussions regarding dimensionality and item stability will be presented in the final paper.



Median dimensionality results for 500 bootstrap ega of the chiropractic osce.png

Using a Bi-Factor Version of MGGUM for Multidimensional Proximity-based Data

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Zhaoyu Wang (Georgia Institute of Technology), Prof. James Roberts (Georgia Institute of Technology)

In the application of traditional methodologies such as PCA and FA to the proximity-based data, misleading information about dimensionality arises due to the linear relationship assumed between the observed data and latent variables when the nature of the data is indicative of a non-linear relationship. This leads to the extra factor phenomena when analyzing unidimensional proximity-based data with linear models (Van Schuur & Kiers, 1994). We propose integrating the bi-factor model with Multidimensional Generalized Graded Unfolding Model (MGGUM) to exam and confirm the dimensionality underlying proximity-based data. The incorporation of a bi-factor model is designed to distinguish variance from both a general and multiple specific factors in an unfolding model framework. Specifically, we will first conduct an exploratory investigation using detrended correspondence analysis which can roughly estimate the number of dimensions in the data and provide starting parameter values to a subsequent MCMC calibration of a bi-factor MGGUM. We will leverage model and item fit statistics such as Q3 which not only diagnose the fit of the model but will also help to assess the fit of lower dimensional models in which particular specific factors are removed (Yen, 1984, 1993). We will use R to prepare data and JAGS to fit the bi-factor MGGUM to newly acquired data (N=800) from a political attitude study. We expect to enhance our understanding of both the structural relationships among questionnaire items and latent variables as well as the number of latent dimensions required to reasonably fit data following from a proximity-based response process.

Item Classification by Functional Principal Component Clustering and Neural Networks

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

James Zoucha (University of Northern Colorado), Dr. Igor Himelfarb (National Board of Chiropractic Examineers), Nai-En Tang (NBCE)

Maintaining structure across forms of a test is important for properly and fairly categorizing examinees into pass or fail classes. This paper presented a practical procedure for classifying items by difficulty levels. The methodology uses functional data analysis (FDA) to cluster one administration's item characteristic curves (ICC) into difficulty groups based on their functional principal components (FPC), then trains a neural network for the prediction of unseen ICCs. Sigmoid curves of items are often similar in shape which can make the categorization of items in difficulty groups challenging. Thus, the utility of this method is that it provides an empirical and consistent process for categorizing items as opposed to categorizing through visualization. Findings of the clustering method revealed almost all discrepancies of difficulty categorization between visualization and FDA differed by only one adjacent difficulty level. Of those discrepancies, the FDA method placed items from the medium to hard range in higher categories about 67% of the time, while the remaining third of discrepancies were very easy to easy items being classified into lower categories compared to the visualization method. The trained neural network had an accuracy of 79.6% with misclassifications only differing by one adjacent class compared to how items were clustered by FDA. This analysis highlights a streamlined method of classifying items whose practicality would be emphasized further in online testing settings where results may come in larger quantities and at varying points throughout each year.

Video-administered questionnaire: Psychometric properties and comparison with a text-based format

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Adam Strojil (Masaryk University), Dr. Hynek Cígler (Masaryk University)

Short videos have become popular online content, especially among young adults. Moreover, most respondents use mobile devices with relatively small displays in online questionnaires, worsening the possibility of reading longer items and questions.

This study examines the feasibility of administering an online psychological questionnaire with items presented as short video recordings of an interviewer (VQ). We compare this format to a traditional text-based administration (TQ) using a between-subject design (*N* = 321). VQ was administered either by a male or a female using two different psychological inventories.

Strict (residual) measurement invariance was established in both employed psychological scales; moreover, even population hyperparameters (latent means and variances) did not differ across conditions. No significant differences were observed in the methods' drop-off rate, reliability, or criterion validity.

We found no evidence that the social desirability or the interviewer had any notable influence on VQ. On the other hand, we observed significant but mixed differences in reaction times. One of the interviewers in VQ led to a faster completion time than the TQ, while the other was slower, suggesting that VQ is not necessarily more time-consuming than traditional online questionnaires.

Finally, respondents rated VQ as the more enjoyable method. It appears that VQ may be a valid, reliable, and potentially more engaging alternative method of questionnaire administration, even though our findings require replication. The possibilities and difficulties of using VQ to collect data based on AI's ability to generate videos will be discussed.

Conducting specification search for partial invariance models with unbalanced data

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Po-Yi Chen (Department of Educational Psychology and Counseling, National Taiwan Normal University)

Introduction. When testing factorial invariance, researchers often use modification indices (MFI) to examine partial invariance. However, a past study has shown that when data are unbalanced (i.e., unequal group sample sizes), the power of chi-square statistics of multiple group confirmatory factor analysis will be affected. Considering MFIs are estimates of the changes in chi-square statistics, it is reasonable to assume its efficacy to detect non-invariance will also be affected. In contrast, the method proposed by Jung & Yoon (2016) based on the Wald tests of the differences between corresponding parameters may be more robust. **Method.** We generated data from two-group measurement models (six five-point indicators per group) and manipulated sample sizes (500, 1000, 1500), patterns of loading non-invariance (x4), and group sample size ratio (1:1 and 1:4). **Results:** The results indicate that although both approaches can effectively control the type I error rates in invariant conditions (≤ 0.1), the perfect recovery rates (PPR) of the Wald test procedure are generally higher than MFI. Furthermore, its PPRs are also more robust to unbalanced data (e.g., the means of decrease in PPRs due to the unbalanced data of the two approaches across conditions are 0.139 & 0.209, respectively. Cohen's D = 0.4). **Conclusions**: Our results suggest that the Wald test approach is more robust in examining partial invariance when data are unbalanced.

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 567-584.

A study on restricted HMMs for latent class attribute transitions

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Theren Williams (University of Illinois, Urbana-Champaign), Prof. Steve Culpepper (University of Illinois at Urbana-Champaign), Prof. Yuguo Chen (University of Illinois, Urbana-Champaign)

across most fields, professionals use various Cognitive Diagnostic Models (CDMs) to understand the underlying attribute profiles of a given subject pool. in some settings, subjects may record information over time, such as a pre/post-assessment pairing. in such cases, CDMS development requires that the potential transitions from one profile to another are accounted for. these changes may occur freely or bound to a given set of rules. existing work leverages Hidden Markov Models (HMM) to model the transitions between attribute profiles over time. our model builds on the existing framework, adapting a classical transitions model into a more general latent class model framework. furthermore, we provide and discuss a Markov Chain Monte Carlo (MCMC) simulation study and its potential applications in psychometrics.

Comparing LPA results by missing data methods in various conditions

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Yewon Kim (Ewha womans university), Prof. Youn-Jeng Choi (Ewha womans university), Junseo Kim (Korea University)

There are several missing treatment methods in Latent Profile Analysis (LPA). However, there has been little investigation into how these techniques work under different missing conditions in LPA. This study compared the results of LPA after applying various missing treatment methods, including Listwise Deletion (LD), Multiple Imputation (MI), Full Information Maximum Likelihood (FIML), and Random Forest (RF).

The study utilized the 2018 TALIS teaching practice data from Korea, with the original dataset comprising 2,551 cases with no missing values. The study simulated various missing conditions, with missing rates of 5%, 15%, and 25% under Missing Completely At Random (MCAR), Missing At Random (MAR), Missing Not At Random (MNAR), and mixed mechanisms. LPA was conducted after applying LD, MI, FIML, and RF methods.

Applying MI, FIML, and RF under certain MAR and mixed MAR and MCAR settings revealed the same 7 classes for LPA with original data (see Table 1). However, the number of identified classes varied under different conditions. Additionally, we measured concordance by calculating the number of common classes and the differences in case assignment ratios between LPA results from the original data and the data treated for missing values (see Figure 1). In the figure, a higher dot and a shorter bar indicate greater congruence. This enabled us to determine which missing treatment methods performed better under specific conditions. This study emphasizes the need of carefully selecting treatment methods based on the missing conditions in LPA and developing more effective strategies for handling MNAR and mixed conditions.

Figure 1. Class Concordance with Original Data Model

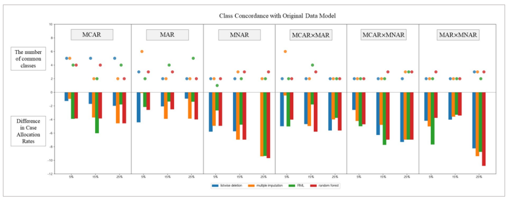


Figure1.png

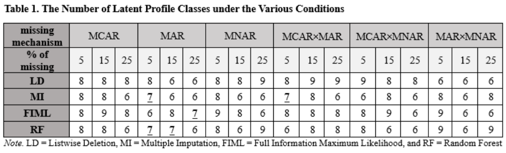


Table 1. The Number of Latent Profile Classes under the Various Conditions

missing mechanism	MCAR			MAR			MNAR			MCAR×MAR			MCAR×MNAR			MAR×MNAR		
% of missing	5	15	25	5	15	25	5	15	25	5	15	25	5	15	25	5	15	25
LD	8	8	8	8	6	6	8	8	9	8	9	9	9	8	8	6	6	6
MI	8	8	6	7	6	6	8	6	6	7	8	6	6	8	6	8	6	8
FIML	8	9	8	6	8	7	9	8	6	8	8	8	8	9	6	9	6	6
RF	8	8	6	7	7	6	8	6	9	6	8	8	8	6	6	9	6	9

Note. LD = Listwise Deletion, MI = Multiple Imputation, FIML = Full Information Maximum Likelihood, and RF = Random Forest

Table1.png

Measurement Invariance of the WJ IV in Clinical Samples

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Hyeonjoo Oh (Riverside Insights), Dr. Tong Wu (Riverside Insights), Dr. Jongpil Kim (Riverside Insights)

To compare test scores meaningfully across different groups, it is crucial that standardized tests yield the same meaning and consistent interpretations across those groups. Given that the standardization of diagnostic assessments is often based on population-representative data, construct validity of such assessments can be significantly affected when conducted on different populations or specific clinical groups. Thus, achieving measurement invariance is an essential requirement to ensure test fairness.

Multigroup confirmatory factor analysis (MGCFA) plays a pivotal role in assessing measurement invariance across diverse subgroups. Nevertheless, research on the measurement invariance of the Woodcock Johnson IV (WJ IV) tests across clinical groups, particularly among children and adolescents with learning disabilities, remains limited and more research on such groups is needed. This study employs MGCFA to examine measurement invariance using standardization data from the WJ IV tests across clinical samples of learning disability (LD) in reading, along with control groups matched for age, gender, race, ethnicity, and region through propensity matching. Preliminary findings of the WJ IV Achievement battery suggest that overall model fits are in an acceptable range (CFI = 0.961, SRMR=0.035, and RMSEA= 0.080) for the unconstrained baseline model indicating configural invariance could be accepted. Additionally, fit indexes for the metric invariance model, constraining all first-order and second-order factor loading to be equal across LD and control groups, also demonstrate acceptable fits (CFI=0.950, SRMR=0.037, and RMSEA=0.087 for first-order metric invariance; and CFI=0.946, SRMR=0.038, and RMSEA=0.089 for second-order metric invariance). More comprehensive results and implications will be discussed during the final presentation.

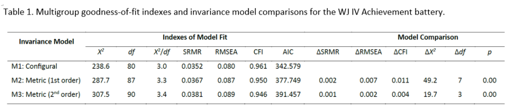
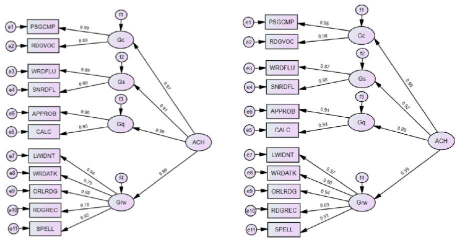


Table 1. Multigroup goodness-of-fit indexes and invariance model comparisons for the WJ IV Achievement battery.

Invariance Model	Indexes of Model Fit							Model Comparison						
	X^2	*df*	X^2/df	SRMR	RMSEA	CFI	AIC	ΔSRMR	ΔRMSEA	ΔCFI	ΔX^2	Δ*df*	*p*
M1: Configural	238.6	80	3.0	0.0352	0.080	0.961	342.579						
M2: Metric (1st order)	287.7	87	3.3	0.0367	0.087	0.950	377.749	0.002	0.007	0.011	49.2	7	0.00
M3: Metric (2nd order)	307.5	90	3.4	0.0381	0.089	0.946	391.457	0.001	0.002	0.004	19.7	3	0.00

Table 1. model fit indexes.png



(a) Control Group

(b) Learning Disability - Reading Group

Figure 1. Standardized regression weights for the WJ IV Achievement battery second-order four-factor model for the LD-Reading and control groups.

Note: PSGCMP = Passage Comprehension, RDGVOC = Reading Vocabulary, WRDFLU = Word Reading Fluency, SNRDFL = Sentence Reading Fluency, APPROB = Applied Problem, CALC = Calculation, LWIDNT = Letter Word Identification, WRDATK = Word Attack, ORLRDG = Oral Reading, RDGREC = Reading Recall, SPELL = Spelling, Gc = Comprehension-Knowledge, Gs = Processing Speed, Gq = Quantitative Knowledge, and Grw = Reading and Writing.

Figure 1. standardized regression weights.png

ChatGPT in the social and health sciences: Can ChatGPT help people improve their health literacy?

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

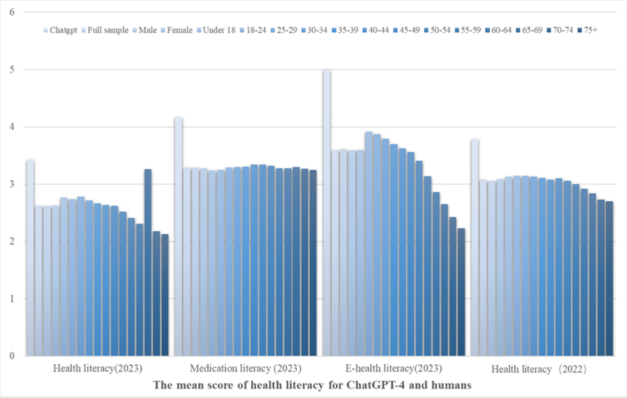
Mr. Shunsen Huang (State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University), Dr. Yibo Wu (School of Public Health, Peking University), Mrs. Huanlei Wang (State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University), Dr. Xiaoxiong Lai (Institute of Digital Education, China National Academy of Educational Sciences), Mrs. Kexin Xiang (State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University), Prof. Cai Zhang (Collaborative Innovation Centre of Assessment for Basic Education Quality, Beijing Normal University), Prof. Fumei Chen (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University), Dr. Li Ke (State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University), Prof. Yun Wang (Faculty of Psychology, Beijing Normal University)

Generative Large Language Models (GLLM, e.g., ChatGPT) have been widely used to guide research in the social and health sciences. Previous studies have assessed the cognitive abilities (Binz & Schulz, 2023), academic performance (OECD, 2023), and personality (Rao et al., 2023) of ChatGPT and compared them to humans. Researchers have advocated the use of ChatGPT to improve people's health literacy or behaviors (Ayre et al., 2023), but there is no evidence whether ChatGPT have better health knowledge or literacy than humans.

Data from two nationally representative datasets (PBICR program collected in 2022(*N*=30505, male=13229, age=11~100) and 2023(*N*=40590, male=17952, age=17~106) (Wang et al., 2022) was used to compare ChatGPT with human health literacy. Health literacy was measured using the Health Literacy Scale, the e-Health Literacy Scale and the Medication Literacy Scale. The ChatGPT (temperature=0.8) was required to respond ten times to each item.

The result showed that health (3.43±0.28 vs 2.63±0.69, p<0.001), e-health (5.0±0 vs 3.60±0.97, p<0.001), and medication (4.17±0.18 vs 3.28±0.59, p<0.001) literacy were significantly higher in ChatGPT than in 2023 samples, even across most single items, gender and age groups. Similarly, health literacy in ChatGPT (3.79±0.10 vs 3.07±0.59, p<0.001) was higher than in 2022 samples, even for most single items, gender and age groups.

ChatGPT has an absolute predominance in health and medication literacy compared to humans, indicating the beneficial prospect of GLLM in guiding human health literacy and behaviors.



Chagpt literacy.png

Fast and efficient distributed Bayesian inference in large-scale educational assessment

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Sainan Xu (Northeast Normal University), Dr. Jing Lu (Northeast Normal University), Dr. Jiwei Zhang (Northeast Normal University), Chun Wang (University of Washington), Dr. Gongjun Xu (University of Michigan)

With the growing attention on large-scale educational testing and assessment, the ability to process substantial volumes of response data becomes crucial. Current estimation methods within item response theory (IRT), despite their high precision, often pose considerable computational burdens with large-scale data, leading to reduced computational speed. This study introduces a novel "divide and conquer" parallel algorithm built on the Wasserstein posterior approximation concept, aiming to enhance computational speed while maintaining accurate parameter estimation. This algorithm enables drawing parameters from segmented data subsets in parallel, followed by an amalgamation of these parameters via Wasserstein posterior approximation. Theoretical support for the algorithm is established through asymptotic optimality under certain regularity assumptions. Practical validation is demonstrated using real-world data from the Programme for International Student Assessment. Ultimately, this research proposes a transformative approach to managing educational big data, offering a scalable, efficient, and precise alternative that promises to redefine traditional practices in educational assessments.

Multivariate Generalizability Theory for Automated Item Generated Test Forms

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Stella Kim (University of North Carolina at Charlotte), Dr. Sungyeun Kim (Incheon National University)

The current study aims to investigate the reliability of test forms generated through automatic item generation (AIG) techniques. With the rising popularity in the use of AIG, particularly in the context of test development facilitated by tools like ChatGPT, there's a growing need for a robust and sound measurement framework to assess the quality of such test forms. However, existing literature lacks a comprehensive framework for evaluating these forms psychometrically.

Generalizability theory (G-theory) has traditionally been a cornerstone for psychometric analyses, especially for data having multiple sources of measurement errors. This study seeks to propose potential analytic designs for AIG-based data using multivariate G-theory, which can offer more nuanced insights into the reliability of these test forms.

Recently, Chung and Kim (2023) analyzed a set of operational data generated by AIG using multivariate G-theory. However, their approach fell short in presenting a comprehensive design suitable for exams with multiple content areas and item formats. Instead, they evaluated content and item format facets separately, which resulted in failing to capture all relevant sources of measurement errors in a single design.

The current study proposes a more holistic design that integrates both content and item format facets, alongside to an item facet, into a unified framework. Also, during the presentation we will delve into a discussion on how to interpret the variability in test forms (consequently in test scores) as a result of personalized test forms generated by AIG within the framework of G-theory.

How many IRT parameters does it take to high stake test?

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Xue Zhang (Northeast Normal University)

Under the unidimensional item response theory framework, 1-, 2-, 3- and 4-parameter models and their variants were proposed by considering different types of item parameters. In practice, the high-stake testing types can be summarized as competitive exams and qualifying exams. For these tests, information of interest was the numerical/scaled values (e.g., SAT), the grades (e.g., A-level) or the ranks (e.g., ATAR, Chinese national college entrance examination) of abilities. For different purposes, how many IRT parameters dose it take? To this end, a pilot study was conducted to illustrate and compare the performances of different UIRT models under different purposes via simulation studies. The MMLE/EM algorithm was used to estimate item parameters, and MLE and EAP algorithms were used to estimate abilities. The root mean squared error (RMSE) and bias were used to evaluate ability recovery, and Wilcoxon signed ranks test, Spearman's rank correlation test and Kendall's W coefficient were used to assess the consistency and uniformity.

The impact attribute hierarchies' distribution on diagnostic classification accuracy

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Chia-ling Hsu (Hong Kong Examinations and Assessment Authority), Dr. Yi-Jhen Wu (The Center for Research on Education and School Development, TU Dortmund)

Cognitive diagnostic models provide detailed information about examinees' mastery of a set of fine-grained discrete latent skills/attributes. This information is useful for researchers and educators to tailor instruction and develop cost-effective interventions to improve students' learning outcomes. A learning process generally assumes that when a student has mastered high-level skill, he/she should first master lower-level skills. Given this assumption, four hierarchical attribute structures, have been introduced (Leighton et al., 2024), including linear, convergent, divergent, and unstructured. However, past research with hierarchical attribute structure relied on the assumption that all students within a sample had the same hierarchal structure. For example, all students have a linear structure in mathematic assessments. This assumption might be too strict and unrealistic because students' learning processes and their characteristics are heterogeneous, which might result in different hierarchical attribute structures for learning. To relax the assumption of a single hierarchical attribute structure within a sample, this study conducted simulations to examine the classification accuracy of cognitive diagnostic models when different attribute hierarchies existed. The simulation results showed that the distribution of hierarchical attribute structures affected the classification accuracy. That is, the greater the heterogeneity of the attribute hierarchies within a sample, the higher classification accuracy was found in a more complex hierarchical attribute structure since it provides more attribute profiles to classify students.

Conditional process analysis with measurement errors via R package silp

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Yi Hsuan Tseng (National Taiwan University), Dr. Po-Hsien Huang (National Chengchi University)

In psychological research, the analysis of mediation and moderation is a critical issue. Conditional Process Analysis (CPA) stands out as a prominent methodology for exploring mediation and moderation, offering a comprehensive framework that seamlessly combines both analytical processes. However, CPA is limited by its dependence on regression models that use only observed variables, which does not address the issue of measurement error attenuation. An alternative approach, the Latent Moderation Structural Equation (LMS), overcomes this limitation but is only available through the commercial software Mplus. To bridge the gap, our research introduces an R package called silp (Single Index Latent Process), which enables CPA implementation that accounts for measurement errors. The silp package allows researchers to flexibly define mediation and moderation effects among latent variables and employs a Reliability-Adjusted Product Indicator (RAPI) method for effect estimation. Our simulations indicate that silp provides estimations of comparable quality to LMS for sufficiently large sample sizes (e.g., N = 500), though its estimates may be less stable with smaller samples (e.g., N = 200). Additionally, we suggest several strategies to enhance silp's performance in small-sample scenarios. Conclusively, silp offers a valuable solution for researchers interested in investigating complex mediation and moderation relations without relying on commercial software.

Keywords: structural equation modeling, mediation, moderation, conditional process analysis

Statistical properties of matrix decomposition factor analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Yoshikazu Terada (Osaka University / RIKEN AIP)

For factor analysis, many estimators, starting with the maximum likelihood estimator, are developed, and the statistical properties of most estimators are well discussed. In the early 2000s, a new estimator was developed based on a matrix factorization for factor analysis called Matrix Decomposition Factor Analysis (MDFA). Although the estimator is obtained by minimizing the principal component analysis-like loss function, this estimator empirically behaves like other consistent estimators of factor analysis, not the principal component analysis. In this talk, to explain this unexpected behavior theoretically, we will show several fundamental statistical properties of MDFA. From these results, we can conclude that the MDFA estimator is appropriate for factor analysis.

Combining results from a large number of cluster analyses: a proof-of-concept analysis

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Anikó Lovik (Leiden University)

In some situations, combining clusters from multiple cluster analyses is necessary. For example, when multiple imputation is applied to missing data or in case of dealing with multilevel data be applying multiple outputation (in settings where the within-cluster correlation is not of interest). In such cases, often tens or hundreds of results must be combined into one final solution.

Combining clusters is far from straightforward. Potential problems include: 1) potentially differing number of clusters in different analyses, 2) the difficulty of matching clusters, for example, based on cluster centroid), 3) deciding on how to calculate cluster centroids of the pooled results, 4) observations potentially changing cluster membership and 5) subsequent difficulty in calculating cluster size.

Here we present a proof-of-concept analysis on a clinical trial dataset of 379 patients of the with rheumatoid arthritis participating in the CareRA trial with data collected at 8 visits. Since most patients were expected to be stable due to treatment, it was assumed that a pooled cluster analysis would be the most useful in separating patients with adverse disease trajectories from those who were responding well to treatment. Hierarchical k-means clustering was combined from 2000 datasets obtained by taking 100 independent samples of each of 20 multiply imputed datasets.

Estimating Controllability Metrics for Ordinal Vector Autoregressive Models

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Bing Cai Kok (University of North Carolina at Chapel Hill), Dr. Kathleen Gates (University of North Carolina at Chapel Hill)

The quantification of symptom importance in psychological disorders is a central problem in clinical science. Many diverse methods have been proposed to tackle this problem, and one such technique is through the application of control theory to psychological time series. In this approach, the evolution of multiple symptoms across time is treated as a dynamical system and symptom importance is quantified through the controllability gramian of the underlying system. Existing work, however, assumes that individual symptoms are measured on continuous scales. This is in contrast to the inherent ordinal nature of most psychological measures. In this regard, it is largely unknown how effectively we can recover the true underlying gramian when ordinal measures are of interest, especially in small sample regimes.

We study the recoverability of the controllability gramian obtained through estimating ordinal vector autoregressive models in an SEM framework. The ordinal nature of the data is handled through polychoric correlations, whereas the autoregressive effects are modeled using the block-toeplitz method paired with diagonally weighted least squares estimation. Across different model configuration settings, simulations reveal that the quality of symptom importance estimates degrade rapidly with decreasing number of ordinal categories in the measurements. These estimates are especially poor when the number of time samples is limited (e.g. < 100 timepoints). In view of these problems, we discuss how the concept of meta-learning, a technique originating from computer science, could potentially alleviate these issues by allowing us to coherently pool information from across multiple individuals when estimating personalized dynamic psychological models.

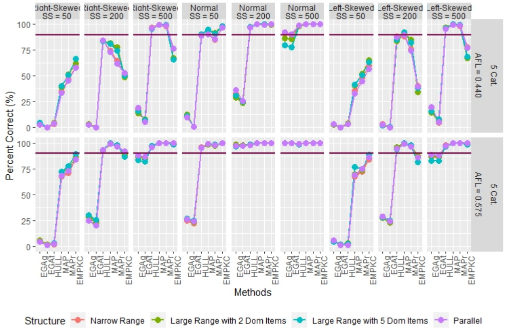
To What Extent Dominant Items Effect on Factor Retention Methods' Performance?

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

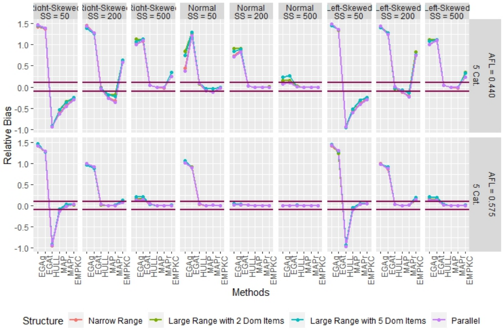
Mr. Tugay Kaçak (Trakya University), Dr. Abdullah Faruk Kılıç (Trakya University), Prof. Meltem ACAR GÜVENDİR (Trakya University), Dr. Gül Güler (Trakya University)

In this study, Zwick & Velicer's (1986) finding that there will be at least 2 dominant items, which means high loadings on a factor, was analysed and simulated. In a large number of Monte Carlo simulation studies, factors were set in a narrow range. For example, conditions with an average factor loading of 0.40, factor loadings are uniformly distributed to represent congeneric measurements, with factor loadings between 0.35-0.45. However, in empirical studies, generally, the factor loadings do not distribute like this setting. The fact that the factor is not saturated is also seen as a problem, and it is a practical problem whether the factor can be reproduced or not. According to this, we focused on two dominant items, five dominant items, items with a narrow range of factor loadings, and items with equal factor loadings. Sample size (50, 200, 500), skewness (-2.5, 0, +2.5), and average factor loading (0.440 and 0.575) were the other conditions, and the test length was 10, the number of categories was 5, and the unidimensional structures were fixed. Totally, 72 conditions and 1000 replications were evaluated using relative bias and percent correct.

Several conclusions can be drawn about determining the number of factors: EGA-TMFG and EGA-Glasso estimated higher than real number of factors, and MAP/MAP4 performed acceptable biased even when sample sizes were small. Zwick & Velicer's suggestion would be more applicable for small and medium sample sizes when the factor is saturated. With this, EGAs and Hull performed oppositely in small samples.



Dty retention uploaded.jpeg



Bias retention uploaded.jpeg

Evaluating DIF within the Vulnerability to Abuse Screening Scale (VASS)

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Michalina Gajdzica (Jagiellonian University Medical College), Katarzyna Zawisza (Jagiellonian University Medical College), Aleksander Galas (Jagiellonian University Medical College), Beata Tobiasz-Adamczyk (Jagiellonian University Medical College), Tomasz Grodzicki (Jagiellonian University Medical College)

The differential item functioning (DIF) analysis is now perceived as especially important, being an essential part of tool development strategy and has been recognized as standard set of techniques to measure significant item function differences across groups while controlling the overall scores on the trait being measured. The study aims to explore differential item functioning by gender of respondents in the Vulnerability to Abuse Screening Scale (VASS). The data used in the analysis comes from a cross-sectional study “Neglect and Self-Neglect” conducted in Lesser Poland in 2017. There were 2001 community-dwelling individuals aged 65+ years randomly selected for the study from among the general population. We reviewed available published data to identify tools of DIF assessment and selected the most popular tools available to identify such items and compare the results of selected techniques. For the purpose of the study the generalized Mantel-Haenszel procedure, the Breslow-Day test, the logistic regression approach and finally IRT mixed effects models were performed to detect uniform and nonuniform DIF. Statistical analysis mainly was performed using SAS procedures and R package’s such *difR* and *mirt*. Some differences in detecting both uniform and nonuniform DIF were found between analyzed techniques.

Anonymisation of data for open science in psychology

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Jiří Novák (University of Zurich), Prof. Carolin Strobl (University of Zurich), Prof. Matthias Templ (University of Applied Sciences and Arts Northwestern Switzerland)

There is a great demand for making more research data openly available. The reproducibility of findings in psychology has been questioned, and more openly available data would make research more transparent and accessible. Unfortunately, many datasets cannot be publicly available for privacy reasons. On the other hand, researchers are increasingly more expected to share data with others for review, reanalysis, and reuse. To solve this issue, we suggest using methods of Statistical Disclosure Control for data anonymisation. These methods either modify or synthesise data so that it can be disclosed without revealing confidential information that may be associated with specific respondents. In this presentation, we review the work in this area and present different anonymisation approaches that can be used to protect data confidentiality. To prove the success of data anonymisation, data utility is discussed as the main objective to be maximised while providing data with a disclosure risk below certain limits. The concepts are illustrated by means of a practical application example.

Exploring Child Well-Being Trends through Psychometric Meta-Analysis of International Large-Scale Assessments

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mrs. Kaitlin Griffith (Charles University)

This study delves into the realms of child well-being and academic achievement through the lens of psychometric meta-analysis applied to data sourced from International Large-Scale Assessments (ILSAs). Despite being in its preliminary stages, this research aims to establish a framework for understanding the intricate dynamics between child well-being indicators and academic performance across diverse cultural and socio-economic contexts.

At this stage, the groundwork involves developing robust search strings tailored to capture relevant data within ILSAs. While comprehensive data analysis may not be attainable by the time of the conference, this study presents a theoretical framework for investigating potential trends and correlations between child well-being metrics and academic achievement outcomes.

The theoretical underpinnings of this research draw from a synthesis of psychometric principles and socio-ecological frameworks, aiming to unravel the multifaceted interplay between individual, familial, and societal factors influencing child well-being and educational attainment.

By harnessing the power of meta-analysis techniques, this study aspires to contribute valuable insights into the nuanced dimensions of child development and educational outcomes on a global scale. The proposed research holds promise for informing policy interventions and educational practices geared towards fostering holistic child development and enhancing academic success across diverse educational contexts.

Opening a Pandora's box of SEM: a systematic review of model characteristics influencing model fit

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Ms. Edita Chvojka (Utrecht University), Petr Palíšek (Masaryk University), Dr. Beth Grandfield (Utrecht University), Prof. Rens van de Schoot (Utrecht University)

Researchers often use Structural Equation Modeling (SEM) to assess the quality of measurement instruments and analyze complex multivariate relationships. Model fit evaluation is a crucial ingredient of SEM. Many researchers agree that no single test indicates that an SEM model fits well. Instead, we can gauge how well a particular model fits the observed data through the collective use of multiple fit indices, such as RMSEA, CFI, TLI and SRMR. However, beyond misspecification, fit indices are also sensitive to additional model characteristics, which is not reflected in any commonly used guidelines. Methodologists have long focused on the sensitivity of fit indices to model misspecification and other characteristics. However, their findings have had a near-zero impact on applied literature. We used ASReview to systematically search literature on the sensitivity of goodness-of-fit indices to different model characteristics. Our poster presentation will introduce the preliminary results of our review. We will provide an overview of the characteristics researched up to date and their effects on different fit indices. Additionally, we will discuss potential reasons for this gap between methodological and applied research and provide an overview of possible solutions that aim to mend this divide.

Assessing Discriminant Validity: Bridging Traditional and Modern Approaches

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Anna Panzeri (University of Padua), Prof. Andrea Spoto (University of Padua)

Discriminant validity (DV) is present when two measures of similar, yet distinct constructs exhibit a correlation that is sufficiently low to consider them as distinct.

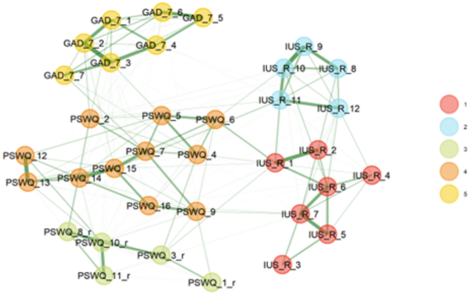
Various techniques can assess DV, recent guidelines suggest that Structural Equation Modeling (SEM) yields the most reliable results, although modern approaches like Exploratory Graph Analysis (EGA) provide valuable insights.

Here we offer a comparison between SEM and EGA methods for assessing DV with an application to measures for which DV has not been previously examined – Penn State Worry Questionnaire (PSWQ); Intolerance of Uncertainty Scale-Revised (IUS); General Anxiety Disorder 7-items (GAD-7)– all part of the 'Digital Intervention in Psychiatric and Psychological Services' (DIPPS) project.

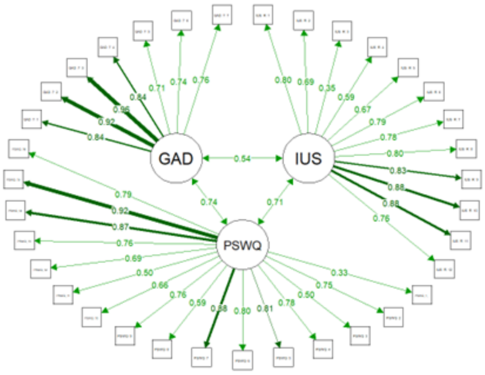
In a SEM incorporating the three measurement models (RMSEA= .053, SRMR= .056; CFI= .991), the latent correlations with 95% bootstrapped CIs were freely estimated and fell below the recommended threshold (<0.85), thus supporting the DV presence (PSWQ~~IUS-R= 0.706, 95%CI[0.695, 0.717]; PSWQ~~GAD-7= .736, 95%CI[.723, .748]; IUS~~GAD-7= 0.540, 95%CI[0.525, 0.554]).

The EGA with parametric bootstrap^{1000} on the items of the three measures correctly identified the constructs' dimensions, with optimal stability indices for items and dimensions (all>.92). EGA loadings' analysis revealed no problematic cross-loadings, supporting the DV among the three measures.

In conclusion, a modern technique such as EGA was applied for the first time in DV assessment providing results consistent with the traditional SEM approach, possibly indicating similarities between the methodologies. Furthermore, as these approaches sometimes produce equally satisfactory results, modern methods could serve as a valuable alternative when traditional methods encounter difficulties.



Imps boot ega plot.png



Imps semplot.png

A LASSO approach for short form item selection

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Klint Kanopka (New York University), Dr. Daphna Harel (New York University)

Constructing shorter versions of surveys, screeners, and assessments can be desirable in clinical and operational settings where the negative impacts of response burden and increased cognitive load can be significant sources of measurement error. Selecting items for these short forms is especially common when dealing with patient-reported outcome measures, health screeners, and the educational assessment of young children. However, the shortening of instruments can threaten reliability and validity (Smith, McCarthy, & Anderson, 2000). Typical methods for item selection include item parameter or information-based approaches derived from Item Response Theory (IRT), Optimal Test Assembly (OTA; Harel & Baron, 2018), and machine learning approaches (Lu & Petkova, 2013; Gonzalez, 2020). We propose a machine learning-based approach that utilizes the Least Absolute Shrinkage and Selection Operator (LASSO) and cross-validation to select items. This process keeps validity in mind by generating an intermediate outcome for prediction from the full-length form aligned with the proposed use of the short form, typically a scale score or classification decision. Using cross-validation, we focus on the development of forms that generalize to new respondent samples. This method allows users to balance induced measurement error against the number of items retained by tuning the regularization parameter. Using simulation, we compare our proposed LASSO-based method with information-based selection and OTA methods across various pilot respondent sample sizes, item pool sizes, number of items retained, and data-generating processes. Additionally, we highlight some situations where our method selects items with higher true test information.

Linguistic features usability analysis in educational context

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Anna Winklerová (Masaryk University)

The ongoing intensive research on linguistic features (LF) engineering and their utilization in machine learning (ML) models in various scenarios is resulting in hundreds of features in lexico-semantic, syntactic and other categories. These amounts of handcrafted and task-specific LF accompanied with different implementations across research works (B. Lee, J. Lee: LFTK-Handcrafted Features in Computational Linguistics, 2023) emerge the need to comprehensively analyze task specific feature selection and model evaluation methods and implement tools for developers to make a reasoned decision in specific applications.

This presentation gives (1) a detailed process outline of the general pipeline for LF utilization and model evaluation (2) together with two specific use cases in both the computerized adaptive testing (CAT) environment and adaptive learning system (ALS) – areas, which are closely related, but mostly studied separately. The domain of the data is second language acquisition (L2 English). The items are multiple choice types which are quite challenging to analyze due to their short textual part. Data in both datasets count in thousands items categorized in multilayer granularity and are answered by thousands of students making this dataset highly suitable for deep step by step analysis of LF usability in series of applications (ie. difficulty prediction).

Outcome of this study demonstrates methods for comprehensive interpretation of pipeline steps such as feature selection, dimension reduction or parameter setting for supervised learning algorithms. These tools aim to support developer's decisions in the process of generating and maintaining high-quality content, modeling student's behavior and/or deploying new system functionality.

Does deleting biased items make selection fairer and more accurate?

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Meltem Ozcan (University of Southern California), Dr. Hok Chio (Mark) Lai (University of Southern California)

Valid and meaningful comparison between factor scores across groups or conditions necessitates measurement invariance (MI), which is achieved when latent construct(s) are measured equivalently and comparably across demographic groups (e.g., ethnicity, SES), test modes (e.g., paper, computer), organizational characteristics (e.g., position, cohort), or time points. Partial measurement invariance (PMI) exists when MI holds for only a subset of items, and the noninvariant item parameters are allowed to vary across groups. Millsap and Kwok (2004) developed a single-factor selection accuracy framework for evaluating the impact of PMI, which Lai and Zhang (2022) have extended to multi-factor tests. In this study, we explored a number of Cohen's *h* effect size indices to quantify the impact of deleting an item on various selection indices (e.g., proportion selected, sensitivity and specificity) for a multi-factor test. We developed an R package and R Shiny web application that allows researchers to examine the improvement or reduction in (a) the fairness and (b) performance of a test for selection purposes under various item deletion scenarios. The software requires inputs of only the CFA parameter estimates, the number of factors, and the number of items in each factor. We use parameter estimates from a previous invariance study involving the Center of Epidemiological Studies Depression (CES-D) Scale to illustrate how the item deletion indices can be used, and provide suggestions for interpretation to help researchers make more informed decisions while considering deleting noninvariant items.

Predicting item difficulty with text analysis and machine learning in different languages and item types

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Filip Martinek (Institute of Computer Science, Czech Academy of Sciences), Jan Netík (Institute of Computer Science, Czech Academy of Sciences; Faculty of Education, Charles University), Dr. Patrícia Martinková (Institute of Computer Science, Czech Academy of Sciences; Faculty of Education, Charles University)

In standardized testing, predicting item difficulty from item wording is useful both for test development as well as for deeper understanding of what makes an item a difficult one. Many features may influence item difficulty, such as the length of answer choices, their similarity with the item question, difficulty of the words used, etc., and different machine learning models may be used to predict item difficulty from item features (Štěpánek et al., 2023). However, differences and challenges may arise when building models for different item types (including those involving audio, or visual components), and for different languages.

In this work, we extract item features from various types of test items from the English, German, and French as foreign languages Czech matura exams into various item features, and train numerous different machine learning models to predict their difficulty. We compare and analyze the models and features in order to create a tool that can analyze and suggest changes during test development to help achieve an optimal item difficulty.

Adaptation and Preliminary Reliability Studies of Basic Interest Markers (BIM)

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Prof. Victor Ortuño (Universidad de la República), Ms. Julieta Cabrera (Universidad de la República), Mr. Diego Capuccio (Universidad de la República), Ms. Micaela Mesa (Universidad de la República)

Vocational interests consist of a set of personal traits that express preference for certain activities, outcomes, and contexts (Rounds & Su, 2014). These represent an intermediate level between specific occupations and abstract dimensions such as general interests, widely used in models like Holland's RIASEC (Day & Rounds, 1997). The Basic Interest Markers (BIM, Liao et al., 2008) consist of an inventory of 337 items (5-point Likert scale) grouped into 31 dimensions that allow evaluating the structure of individual interests under a more comprehensive model than that presented by Holland. This work serves to present the process of translation and cultural adaptation of the BIMs, as well as the first results of test-retest reliability and internal consistency.

The sample consists of 41 Psychology students, aged between 21 and 51 years (M = 28.2, SD = 6.9). 26 (66.7%) of these are female and 13 (33.3%) are male. Three independent translations of the instrument were performed, then discussed in an expert panel and used at two moment with approximately 20 days apart.

The overall values of test-retest reliability ranged from moderate to high (rTime1-Time2 = .32, .94), as did internal consistency values (α = .85, .98) across de 31 dimensions.

The adaptation of the BIMs to Uruguay has shown very adequate levels of test-retest reliability and internal consistency, similar to those found in the original version of the instrument. Future studies are suggested to explore the factorial structure and other validity criteria of the instrument.

Beyond Normal: Exploring RMSD Performance with Alternative Theta Distributions

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Mr. Preston Botter (Indiana University), Dubravka Svetina (Indiana University), Dr. Sijia Huang (Indiana University)

The root mean square deviation (RMSD) is a popular tool for detecting differential item functioning (DIF) and has been routinely applied operationally in international large-scale assessments (ILSAs) such as PISA. Joo et al. (2023) proposed several promising weights to the RMSD statistic to improve its power to detect DIF in low performing countries. The aim of the present study is two-fold. First, we evaluate the performance of proposed RMSD alternatives when the latent ability (i.e., theta) distribution is non-normal. Second, we examine the impacts of parameterizations of the theta distribution on the performance of RMSD. Specifically, we apply these weights to an alternative RMSD formulation that allows for non-normal representations of theta distribution. We estimate the theta distribution using both the empirical histogram (EH; Bock & Aitkin, 1981; Woods, 2007) method and the skew-normal distribution and evaluate their performance using an extensive simulation study with conditions like that of Joo et al. (2023). Our study contributes to both the knowledge of detecting DIF in the ILSA context and fitting IRT models to data with many groups using non-normal theta parameterizations. To our knowledge, several studies have investigated the EH method in the case of one group and recommended freeing upwards of 100 additional parameters when fitting an EH IRT model. However, no study has evaluated whether the EH method can be used for 10 or more groups, such as 50+ groups as in the case of ILSAs. Preliminary results focusing only on the EH method found that it shows promise.

Impact of priors on parameter estimates of the Rasch model

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Jordan Wheeler (University of Nebraska-Lincoln), Sarah Hammami (University of Nebraska-Lincoln), Elia Harper (University of Nebraska-Lincoln), Paul Hermanto (University of Nebraska-Lincoln), Sunhyoung Lee (University of Nebraska-Lincoln), Jamy Rentschler (University of Nebraska-Lincoln), Garrett Wirka (University of Nebraska-Lincoln)

This study explored the effects of prior distributions and hyperparameters on the estimation of parameters for the Rasch model through a simulation using various conditions. Specifically, the study investigated the effectiveness of using data-driven, informative priors for the estimation of person abilities and item difficulties. The normal distribution was used as the prior distributions. The centers of the prior distributions were chosen using the quantile function of the normal distribution where the inputs were chosen based on the percentage of items a person answered correctly for the prior distribution of person ability, and the proportion of people who answered the item incorrectly for the prior distribution of item difficulty. In addition to these priors, four factors were manipulated and all factors were crossed for a total of 500 conditions. The four factors include: sample size (4 levels: 50, 100, 250, 500), number of items (5 levels: 10, 20, 30, 40, 50), the center of the normal distributions for generating person ability (5 levels: -3, -1, 0, 1, 3; variances = 1), and the center of the normal distributions for generating item difficulty (5 levels: -3, -1, 0, 1, 3; variances = 1). The results of the simulation showed that using informative priors performed similarly to less informative priors (i.e., standard normal) when person abilities and item difficulties were generated from a standard normal distribution. However, the informative priors performed better than the less informative priors when person abilities and item difficulties were generated from off-centered distributions with little overlap.

The performance of different shrinkage parameter decision principle for GME-EFA

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Yen Lee (Uniformed Service University)

Lee (2020) proposed a generalized maximum estimator for EFA (GME-EFA), a type of shrinkage estimator that circumvents Heywood cases and produces more solutions which recover the correct factor structure when the sample size is relatively small and the data are ordinal. For shrinkage estimators, selecting a shrinkage parameter value is crucial to obtaining high-quality solutions. In this study, we examined the performance of two criteria suggested to determine the shrinkage parameter value (Jin et al. 2018): the sum of the differences and the Kullback-Leibler (K-L). Both criteria measure the deviation between the reproduced correlation matrix and the data correlation matrix. To evaluate their performance in selecting the shrinkage parameter value, a large scale simulation study with six factors was conducted: (a) model complexity, (b) factor loadings, (c) sample size, (d) variable measure and distribution, (e) factor correlation, and (f) shrinkage parameter value determination criteria. The mean squared errors, the Tucker's congruence coefficients, the proportion of solutions that recovers the correct factor structure of each condition was examined. The performance of the criteria were discussed, and several recommendations were made.

The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis: A replication

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Jeremy Miles (Google), Alexander Miles (University of Southern California), Mark Shevlin (University of Ulster)

There have been recent calls for researchers in social science methodological re-search to consider replication as an This paper reports on a replication of Curran, West, & Finch (1996), a simulation study that used EQS 3.0 to generate random data and analyze it in a confirmatory factor analysis framework. We present a replication of this simulation using more recently developed, open source software(the simsem and lavaan packages in R). The results that we obtain are substantively equivalent to the results obtained in the original paper, but some minor discrepancies were found, and we discuss the possible reasons. We conclude with an argument that replication of simulation studies can be useful and informative, and thanks to the rise of open source analysis software, websites that increase ability to share code and improvements in computer hardware, the costs of replication are dramatically reduced.

The issues with alternative fit indices of CFI and RMSEA

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

WENJUO LO (University of Arkansas)

Confirmatory factor analysis (CFA) has been extensively applied to examine measurement invariance across multiple populations. However, the major issue of chi-square test, commonly used in CFA, is sensitive to sample size. As a result, alternative fit indices (AFIs), such as the change in Comparative Fit Index (ΔCFI) and root mean square error of approximation (ΔRMSEA), have been recommended to detect non-invariant parameters among groups. Despite the utility of these AFIs, discrepancies in evaluations between CFI and RMSEA in initial data-model fit and their extension to other forms of invariance evaluation remain unexplored. This study evaluated these AFIs across three levels of invariance (i.e., configural, metric, and scalar) under various data generation conditions by the number of variables with differing factor loadings between groups, the magnitude of differences in factor loadings between groups, correlations between factors, and sample size. The findings provide insights into the performance of ΔCFI and ΔRMSEA in detecting non-invariant parameters and offer recommendations and discussions on extensions and limitations.

Complementing Rasch-trees with various DIF detection methods for media-addiction scale

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Yewon Kim (Ewha womans university), Dr. Minsun Chi (Seoul National University), Junseo Kim (Korea University), Prof. John Jongho Park (Pillar foundation)

The Rasch trees analysis is a way to detect differential item functioning (DIF) in the Rasch model that combines recursive partitioning. While this method has the advantage of identifying explanatory variables that contribute the most to DIF, there are no clear criteria for determining which item is DIF. Therefore, this study examined whether additional DIF detection techniques could compensate for the results of the Rasch trees analysis.

This study used the data from the Panel data on Korean Children, which involved 1348 mothers' responses to their children's media addiction. The scale consisted of 15 items on a 4-point Likert scale. Thirty-two explanatory variables were considered, including mothers' psychological and parenting-related factors. Next, we conducted IRT-LR, Logistic Regression (LR), and Mantel-Haenszel (MH) to compare to the DIF analysis result from Rasch trees analysis, which was performed according to the significant splitting criterion.

The Rasch trees showed that there were the DIF items between groups with children spending less than 2 hours alone without an adult (TIMEALONE ≤ 3 points), which is an explanatory variable, and those spending 2 hours or more (see Figure 1). When the same criterion was used to detect the DIF by IRT-LR, LR, and MH, items 4, 10, and 11 appeared to consistently emerge as the DIF items, matching the top three items of differences in item difficulty between groups in the Rasch trees (see Table 1). The findings imply that conducting IRT-LR, LR, and MH along with Rasch trees analysis contributes to detecting DIF items.

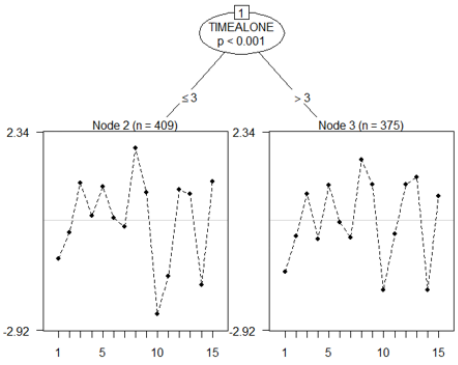


Figure 1. Results of Rasch trees analysis on a media addiction scale

Figure1.png

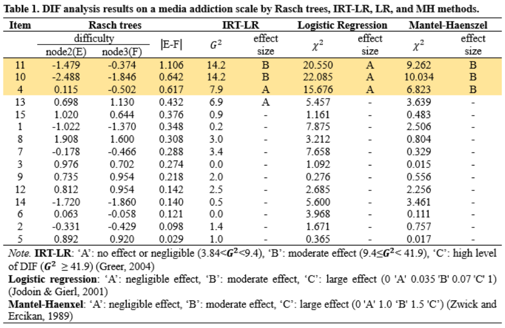


Table 1. DIF analysis results on a media addiction scale by Rasch trees, IRT-LR, LR, and MH methods.

Item	Rasch trees			IRT-LR		Logistic Regression		Mantel-Haenszel	
	difficulty		|E-F|	G^2	effect size	χ^2	effect size	χ^2	effect size
	node2(E)	node3(F)							
11	-1.479	-0.374	1.106	14.2	B	20.550	A	9.262	B
10	-2.488	-1.846	0.642	14.2	B	22.085	A	10.034	B
4	0.115	-0.502	0.617	7.9	A	15.676	A	6.823	B
13	0.698	1.130	0.432	6.9	A	5.457	-	3.639	-
15	1.020	0.644	0.376	0.9	-	1.161	-	0.483	-
1	-1.022	-1.370	0.348	0.2	-	7.875	-	2.506	-
8	1.908	1.600	0.308	3.0	-	3.212	-	0.804	-
7	-0.178	-0.466	0.288	3.4	-	7.658	-	0.329	-
3	0.976	0.702	0.274	0.0	-	1.092	-	0.015	-
9	0.735	0.954	0.218	2.0	-	0.276	-	0.556	-
12	0.812	0.954	0.142	2.5	-	2.685	-	2.256	-
14	-1.720	-1.860	0.140	0.5	-	5.600	-	3.461	-
6	0.063	-0.058	0.121	0.0	-	3.968	-	0.111	-
2	-0.331	-0.429	0.098	1.4	-	1.671	-	0.757	-
5	0.892	0.920	0.029	1.0	-	0.365	-	0.017	-

Note. **IRT-LR**: 'A': no effect or negligible (3.84<G^2<9.4), 'B': moderate effect (9.4≤G^2< 41.9), 'C': high level of DIF (G^2 ≥ 41.9) (Greer, 2004)

Logistic regression: 'A': negligible effect, 'B': moderate effect, 'C': large effect (0 'A' 0.035 'B' 0.07 'C' 1) (Jodoin & Gierl, 2001)

Mantel-Haenxel: 'A': negligible effect, 'B': moderate effect, 'C': large effect (0 'A' 1.0 'B' 1.5 'C') (Zwick and Ercikan, 1989)

Table1.png

Adaptation and Preliminary Validity Results of Questionnaire on Negative Stereotypes towards Old Age (CENVE)

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Prof. Carolina Guidotti (Universidad de la República), Prof. Victor Ortuño (Universidad de la República)

The term "ageism" (Butler, 1968) designates those negative stereotypes and discriminations that are applied to older people simply based on their age. Stereotypes can be defined as those beliefs about certain traits that are supposed to be typical or characteristic of certain social groups, based on ambiguous and incomplete information (Vinacke, 1956).

The Questionnaire on Negative Stereotypes towards Old Age (CENVE, Blanca et al., 2005) consist of an inventory of 21 items (5-point Likert scale) grouped into 3 factors (health, motivational-social and personality) that assess individual negative stereotypes and prejudices towards old populations. This work serves to present the process of translation and cultural adaptation of CENVE, as well its first results regarding its criterion validity.

The sample consists of 440 people aged between 18 and 74 years old (M = 31.6, SD = 11.3). 385 (87.7%) of these are female and 54 (12.3%) are male. CENVE global score was associated with Questionnaire of attitudes towards old age (CAV, Hernandez-Pozo, 2009).

The results were moderate to strong correlations between CENVE and Fear of own aging ($r = .52$, $p < .01$), Negative physical stereotypes ($r = .35$, $p < .01$) and Fear of cognitive impairment and abandonment ($r = .51$, $p < .01$).

The adaptation of CENVE to Uruguay has shown very adequate levels of criterion validity. Future studies should explore its factor structure and reliability.

Evaluating latent mean differences across studies with meta-analytic CFA

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Suzanne Jak (University of Amsterdam)

Meta-analytic Structural Equation Modeling (MASEM) combines the strengths of meta-analysis and structural equation modeling. First, covariance (or correlation) matrices from different studies are combined to form a pooled covariance matrix in a multivariate meta-analysis. Then, a structural equation model is fitted to the pooled covariance matrix. Two-stage SEM consists of these two stages, while one-stage MASEM immediately restricts the pooled covariances from the multivariate meta-analysis to a SEM model.

Currently, there exist no methods for evaluating differences in observed or latent means in MASEM, even though meta-analytic CFA is applied frequently. Examples are the evaluation of instruments for alexithymia, neuropsychological status, or implicit theories of intelligence.

I present and illustrate a method to incorporate the means of variables in meta-analytic structural equation modeling analyses. Meta-analytic CFA with means is applicable when the studies included in the meta-analysis used the same items, measured on the same scales. Applying meta-analytic CFA with means then allows testing differences in latent means across (subgroups of) studies. The new model consists of separate meta-analytic models for the covariance and mean structures, with equality constraints on the parameters that feature in both the model for the means and the model for the covariances. I will illustrate the method in OpenMx using data obtained from 50 studies applying the International Trauma Questionnaire, comparing the latent means of the common factors across samples from the clinical population versus samples from the general population.

A workflow for preprocessing measurements of movement

Wednesday, 17th July - 16:45: Poster session (Atrium RB) - Poster

Dr. Niels Vanhasbroeck (University of Amsterdam), Dr. Tessa Blanken (University of Amsterdam), Prof. Denny Borsboom (University of Amsterdam), Dr. Dora Matzke (University of Amsterdam), Prof. Andrew Heathcote (University of Amsterdam)

Recent technological advances have facilitated the collection of movement data, allowing for a wide range of applications, from analyzing how soccer players move around the field to modeling the pedestrian flow during an evacuation. To measure these kinds of movements, researchers increasingly use local ultra-wide band positioning systems. These systems measure the real-time positions of a "tag" through the signal that it emits. This signal is picked up by "anchors" that are set up at the corners of the measured space and then trilateration is used to determine the tag's location, thus providing the researcher with a coordinate. While these systems are promising, it is not clear to which extent the resulting measurements are subject to systematic and unsystematic error. Additionally, it is not clear which factors might influence the presence of this error and how it can be alleviated. In this talk, I will address these issues and discuss how researchers may be able to separate the signal from the noise.

Authors Index

IMPS 2024

Prague | Czech Republic